On Rates of Convergence for Stochastic Optimization Problems Under Non-I.I.D. Sampling

Tito Homem-de-Mello
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208-3119, U.S.A.
e-mail: tito@northwestern.edu

April 14, 2006

Abstract

In this paper we discuss the issue of solving stochastic optimization problems by means of sample average approximations. Our focus is on rates of convergence of estimators of optimal solutions and optimal values with respect to the sample size. This is a well studied problem in case the samples are independent and identically distributed (i.e., when standard Monte Carlo is used); here, we study the case where that assumption is dropped. Broadly speaking, our results show that, under appropriate assumptions, the rates of convergence for *pointwise estimators* under a sampling scheme carry over to the optimization case, in the sense that convergence of approximating optimal solutions and optimal values to their true counterparts has the same rates as in pointwise estimation.

Our motivation for the study arises from two types of sampling methods that have been widely used in the Statistics literature. One is Latin Hypercube Sampling (LHS), a stratified sampling method originally proposed in the seventies by McKay, Beckman, and Conover (1979). The other is the class of quasi-Monte Carlo (QMC) methods, which have become popular especially after the work of Niederreiter (1992). The advantage of such methods is that they typically yield pointwise estimators which not only have lower variance than standard Monte Carlo but also possess better rates of convergence. Thus, it is important to study the use of these techniques in sampling-based optimization. The novelty of our work arises from the fact that, while there has been some work on the use of variance reduction techniques and QMC methods in stochastic optimization, none of the existing work — to the best of our knowledge — has provided a theoretical study on the effect of these techniques on rates of convergence for the optimization problem. We present numerical results for some two-stage stochastic programs from the literature to illustrate the discussed ideas.

Key words: Stochastic optimization, two-stage stochastic programming with recourse, Monte Carlo simulation, variance reduction techniques, quasi-Monte Carlo methods, Latin Hypercube sampling.

1 Introduction

In this paper we consider stochastic optimization problems of the form

$$\min_{x \in X} \{g(x) := \mathbb{E}[G(x, \boldsymbol{\xi})]\}, \tag{1.1}$$

where X is a subset of $\mathbb{R}^n$, $\boldsymbol{\xi}$ is a random vector in $\mathbb{R}^s$ and $G : \mathbb{R}^n \times \mathbb{R}^s \rightarrowtail \mathbb{R}$ is a real valued measurable function. We refer to (1.1) as the "true" optimization problem. The class of problems falling into the framework of (1.1) is quite large, and includes two-stage stochastic programs as a particular instance.

Oftentimes the expectation in (1.1) cannot be calculated exactly, particularly when G does not have a closed form. In those cases, approximations based on sampling are usually the alternative. One such approximation can be constructed as follows. Consider a family $\{\hat{g}_N(\cdot)\}$ of random approximations of the function $g(\cdot)$, each $\hat{g}_N(\cdot)$ being defined as

$$\hat{g}_N(x) \ := \ \frac{1}{N} \sum_{j=1}^{N} G(x, \boldsymbol{\xi}^j), \tag{1.2}$$

where $\{\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^N\}$ is a sample from the distribution of $\boldsymbol{\xi}$. When $\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^N$ — viewed as random variables — are independent and identically distributed (i.i.d.) the quantity $\hat{g}_N(x)$ is called a *(standard) Monte Carlo* estimator of $g(x)$.

Given the family of estimators $\{\hat{g}_N(\cdot)\}$ defined in (1.2), one can construct the corresponding approximating program

$$\min_{x \in X} \hat{g}_N(x). \tag{1.3}$$

Let $\hat{x}_N$ and $\hat{\nu}_N$ denote respectively an optimal solution and the optimal value of (1.3). Then, $\hat{x}_N$ and $\hat{\nu}_N$ provide approximations respectively to an optimal solution x^* and the optimal value ν^* of the true problem (1.1). Note that the optimization in (1.3) is performed for a fixed sample; for that reason, this is called an *external sampling* approach. When $\hat{g}_N(\cdot)$ is a standard Monte Carlo estimator of $g(\cdot)$, such an approach is found in the literature under the names of sample average approximation method, stochastic counterpart, and sample-path optimization, among others.

The external sampling approach with standard Monte Carlo has been implemented in various settings, see for instance Gürkan, Özge, and Robinson (1999), Kleywegt, Shapiro, and Homem-de-Mello (2001), Plambeck, Fu, Robinson, and Suri (1996). One advantage of that approach lies in its nice convergence properties; for example, it is possible to show that, when x^* is the unique optimal solution, $\hat{x}_N \to x^*$ and $\hat{\nu}_N \to \nu^*$ under fairly general assumptions (see, e.g., Dupačová and Wets, 1988, King and Rockafellar, 1993, Robinson, 1996, Shapiro, 1991, 1993). Two properties have proven particulary useful in terms of establishing *rates of convergence*: the first establish that, under proper conditions, $P(|g(\hat{x}_N) - g(x^*)| \leq \varepsilon)$ and $P(\|\hat{x}_N - x^*\| \leq \varepsilon)$ converge to one *exponentially fast* in the sample size N for any fixed $\varepsilon > 0$ (Dai, Chen, and Birge, 2000, Kaniovski, King, and Wets, 1995). Under some further conditions one can say more, namely, that $P(\hat{x}_N = x^*)$ converges to one exponentially fast in the sample size N (Shapiro and Homem-de-Mello, 2000). Exponential rates of convergence

have interesting consequences in terms of complexity of the underlying problems; see Shapiro (2006) for a discussion.

Another useful property establishes that the sequence of optimal values $\{\hat{\nu}_N\}$ satisfies a certain kind of *Central Limit Theorem* (CLT). More specifically, one has

$$N^{1/2}(\hat{\nu}_N - \nu^*) \xrightarrow{d} \text{Normal}(0, \sigma^*),$$

where "$\xrightarrow{d}$" denotes convergence in distribution and $\sigma^* := \text{Var}[G(x^*)]$ (Shapiro, 1991). An immediate conclusion from the above result is that the rate of convergence of optimal values of (1.3) is of order $N^{-1/2}$. A compilation of these and other related results can be found in Shapiro (2003).

It is no surprise that the sequence of approximating optimal values converges at rate $N^{-1/2}$. Indeed, consider the estimator $\hat{g}_N$ defined in (1.2), and fix $x \in X$. Under mild conditions, it follows from the Central Limit Theorem that $\sqrt{N}[\hat{g}_N(x) - g(x)]/\sigma(x)$ converges in distribution to the standard Normal, where $\sigma^2(x)$ is the variance of $G(x)$. This implies that the error $\hat{g}_N(x) - g(x)$ converges to zero at the rate $N^{-1/2}$. That is, even the pointwise estimators converge at rate $N^{-1/2}$. In many practical cases, the value of N necessary to obtain a reasonable small error under this scheme becomes prohibitively large, especially if evaluation of $G(x, \boldsymbol{\xi})$ for a given $\boldsymbol{\xi}$ is computationally expensive. This motivates the use of *variance reduction techniques* that can yield estimators with smaller variance than the ones obtained with standard sampling. Consequently, the same error can be obtained with less computational effort, which is a crucial step for the use of sampling-based methods in large-scale problems.

Several variance reduction techniques have been developed in the Simulation and Statistics literature, notably importance sampling, control variates, stratified sampling, and others (see, e.g., Bratley, Fox, and Schrage, 1987, Fishman, 1997, Law and Kelton, 2000). However, incorporation of these techniques into a stochastic optimization algorithm is still at an early stage. Existing work (Bailey, Jensen, and Morton, 1999, Dantzig and Glynn, 1990, Emsermann and Simon, 2000, Higle, 1998, Infanger, 1994, Shapiro and Homem-de-Mello, 1998) already shows that significant benefits can be gained by implementing some of these methods, but these papers only provide *empirical* evidence of the gain.

Another approach to obtain better pointwise estimators is to choose the sample points in an appropriate manner. Such is the case of *quasi-Monte Carlo methods* (QMC); see Niederreiter 1992 for a comprehensive discussion, and the brief review we provide in section 3.2. This class of methods has been gaining popularity in the past few years, as it has been observed that these techniques can provide rates of convergence for pointwise estimators superior to the $N^{-1/2}$ obtained with standard Monte Carlo.

A few papers study the optimization problem $\min_{x \in X} \hat{g}_N(x)$ under QMC: Kalagnanam and Diwekar (1997) provide empirical results for the use of Hammersley sequences (one form of QMC) in stochastic optimization problems; Koivu (2005), Pennanen and Koivu (2005) and Pennanen (2005) show that, under mild assumptions, the estimator function $\hat{g}_N$ constructed with QMC points *epiconverges* to the true function g, which guarantees convergence with probability one of optimal values and optimal solutions. Their numerical results also suggest considerable gains in terms of rates of convergence when using QMC methods. Pflug (2004) studies a different type of QMC whereby the sampling points are chosen in a way

to minimize the so-called Wasserstein distance between the original distribution and the empirical distribution generated by the points. Again, the numerical results in Pflug (2004) suggest considerable advantage over standard Monte Carlo.

The above discussion shows that, while there has been some work on the use of variance reduction techniques and QMC methods in stochastic optimization, none of these papers has provided a theoretical study on the effect of these techniques on *rates of convergence*. The reason is that, without the i.i.d. assumption, many of the classical results in probability theory cannot be applied. One exception is the work of Dai et al. (2000), who provide results on exponential rate of convergence of optimal solutions even without the i.i.d. assumption. However, that paper does not focus on any particular sampling technique; rather, they assume that certain conditions that allow for the application of the Gartner-Ellis Theorem in large deviations theory (see, e.g., Dembo and Zeitouni 1998) are satisfied.

In this paper we propose a study of rates of convergence for optimal solutions and optimal values of the approximating problem (1.3) *without imposing that the sample be independent or identically distributed*. Our basic requirement is that $\hat{g}_N(x) \to g(x)$ with probability one for all x, although we shall impose other conditions as we proceed. More specifically, we show that (i) if the proposed sampling scheme yields exponential rate of convergence for *pointwise estimators*, then the convergence of *optimal solutions* will also have an exponential rate, and (ii) if the proposed sampling scheme yields a CLT for *pointwise estimators*, then the convergence of *optimal values* will obey the CLT as well. The setting is fairly general — i.e. the decision space can be continuous or discrete, and the distributions of the underlying random variables can be continuous or discrete, although some the results will not be valid in some of these cases.

We illustrate the ideas for the particular cases of Latin Hypercube Sampling (LHS) and a specific variation of randomized QMC called scrambled (t, m, s)-nets. We show that, for a particular class of functions, the exponential feature of the rate of convergence is preserved under LHS for pointwise estimators and therefore for estimators of optimal solutions. We also use CLT-type results available for LHS and randomized QMC to illustrate the convergence results for estimators of optimal values. In particular, we show that, under LHS, the estimators $\hat{\nu}_N$ of optimal values converge at a rate of order $N^{-1/2}$, the same as standard Monte Carlo; for QMC, under appropriate assumptions the sequence $\{\hat{\nu}_N\}$ converges at a rate of order $[(\log_b N)^{s-1}/N^3]^{1/2}$, which asymptotically is much better than $N^{-1/2}$.

We then apply our results to two-stage stochastic linear programs, and discuss the validity of our assumptions in that context. Numerical results are presented for two problems from the literature to illustrate the ideas.

The remainder of the paper is organized as follows: in section 2 we describe our main results for rates of convergence of estimators of optimal solutions and optimal values. In section 3 we apply these results to Latin Hypercube Sampling and randomized quasi-Monte Carlo. We illustrate the ideas for two-stage stochastic programs in section 4 and present numerical results in section 5. Concluding remarks are presented in section 6.

2 Rates of convergence

We discuss separately the results on rates of convergence for optimal solutions and optimal values. Throughout this paper, S^* and S_N denote the set of optimal solutions of respectively (1.1) and (1.3). Before we study the two cases, we shall make some general assumptions.

Assumption A1: For each $x \in X$, $\hat{g}_N(x) \to g(x)$ with probability one (denoted w.p.1).

Assumption A1 is very natural, as it requires the estimators to be *consistent*. In the i.i.d. case, this is just the standard Strong Law of Large Numbers, which holds if $\mathbb{E}[|\hat{g}_N(x)|] < \infty$ for each $x \in X$.

We make now an assumption on the integrand G viewed as a function of its first argument:

Assumption A2: The feasibility set X is compact and there exists a measurable function $L : \mathbb{R}^s \rightarrowtail \mathbb{R}$ such that $L(\boldsymbol{\xi}) > 0$ w.p.1, $\mathbb{E}[L(\boldsymbol{\xi})] < \infty$ and, for almost every $\boldsymbol{\xi}$ and all $x, y \in X$,

$$|G(x, \boldsymbol{\xi}) - G(y, \boldsymbol{\xi})| \ \leq \ L(\boldsymbol{\xi})\|x - y\|. \tag{2.1}$$

Clearly, Assumption A2 ensures that the function $G(\cdot, \boldsymbol{\xi})$ is continuous for almost every $\boldsymbol{\xi}$. Moreover, it implies that $\hat{g}_N(\cdot)$ and $g(\cdot)$ are also Lipschitz continuous with constants respectively equal to $\hat{L}_N := N^{-1} \sum_{j=1}^{N} L(\boldsymbol{\xi}^j)$ and $\mathbb{E}[L(\boldsymbol{\xi})]$. From (Hiriart-Urruty and Lemarechal, 1993, Theorem IV.3.1.2), we see that if (i) the feasibility set X is compact and contained in the relative interior of the domain of $G(\cdot, \boldsymbol{\xi})$ for almost every $\boldsymbol{\xi}$, and (ii) $G(\cdot, \boldsymbol{\xi})$ is *convex* for almost every $\boldsymbol{\xi}$, then the existence of $L(\boldsymbol{\xi})$ in Assumption A2 is assured, so in that case only finiteness of $\mathbb{E}[L(\boldsymbol{\xi})]$ needs to be checked.

In some cases we will replace Assumption A2 with the following condition:

Assumption A3: Either (i) the feasibility set X is finite, or (ii) X is compact, convex and polyhedral, the function $G(\cdot, \boldsymbol{\xi})$ is piecewise linear for every value of $\boldsymbol{\xi}$, and the distribution of $\boldsymbol{\xi}$ has finite support.

It is worthwhile noticing that, under Assumptions A1 and A2, it is known that (see, e.g., Rubinstein and Shapiro 1993, p.67-70):

(i) $\hat{g}_N(x) \to g(x)$ uniformly on X w.p.1.;

(ii) $\hat{\nu}_N \to \nu^*$ w.p.1;

(iii) $\mathrm{dist}(\hat{x}_N, S^*) \to 0$ w.p.1.

It is not difficult to see that the above holds under Assumptions A1 and A3 as well — in fact, in that case the result in (iii) is replaced with (iii)′: $\hat{x}_N \in S^*$ w.p.1 for N large enough (cf. proof of Theorem 2.1).

2.1 Convergence of approximating solutions

We start by making the following probabilistic assumption on the estimators $\{\hat{g}_N(x)\}$:

Assumption B1: For each $x \in X$, there exist a number $C_x > 0$ and a function $\gamma_x(\cdot)$ such that $\gamma_x(0) = 0$, $\gamma_x(z) > 0$ if $z > 0$, and

$$P\left(|\hat{g}_N(x) - g(x)| \geq \delta\right) \ \leq \ C_x \, e^{-N\gamma_x(\delta)} \qquad \text{for all } N \geq 1 \text{ and all } \delta > 0. \tag{2.2}$$

That is, the probability that the deviation between $\hat{g}_N(x)$ and $g(x)$ is bigger than δ goes to zero exponentially fast with N. Notice that (2.2) implies that $\hat{g}_N(x)$ converges in probability to $g(x)$, which is also ensured by Assumption A1.

Instead of (2.2), we can impose the following weaker condition.

Assumption B1′: For each $x \in X$, there exists a function $\gamma_x(\cdot)$ such that $\gamma_x(0) = 0$, $\gamma_x(z) > 0$ if $z > 0$, and

$$\limsup_{N\to\infty} \frac{1}{N} \log P\left(|\hat{g}_N(x) - g(x)| \geq \delta\right) \;\leq\; -\gamma_x(\delta) \qquad \text{for all } \delta > 0. \tag{2.3}$$

Some of our results will be stated assuming B1 holds; alternatively, B1′ can be used, though in such cases the corresponding result will be stated in asymptotic form as well.

We study now a sufficient condition for Assumption B1 to hold. The main concept behind it arises from the theory of *large deviations*, a well-studied field. For a thorough exposition of the theory, we refer to any of the classical texts in the area, e.g., Dembo and Zeitouni (1998). We present here a result from Drew and Homem-de-Mello (2005).

Proposition 2.1. *Consider the sample $\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^N$ used in* (1.2)*, and define the extended real-valued function*

$$\phi_N(x,t) \;:=\; \frac{1}{N} \log \mathbb{E}\left[e^{tN\hat{g}_N(x)}\right]. \tag{2.4}$$

Suppose that for each $x \in X$ these exists an extended real-valued function ϕ_x^* such that $\phi_N(x,\cdot) \leq \phi_x^*(\cdot)$ for all N, and assume that ϕ_x^* satisfies the following conditions: (i) *$\phi_x^*(0) = 0$;* (ii) *$\phi_x^*(\cdot)$ is continuously differentiable and strictly convex on a neighborhood of zero; and* (iii) *$(\phi_x^*)'(0) = g(x)$. Then, Assumption B1 holds, with the functions $\gamma_x(\cdot)$ given by $\gamma_x(\delta) := \min\{I_x(g(x)+\delta), I_x(g(x)-\delta)\}$ (where $I_x(z) = \sup_{t\in\mathbb{R}}\{tz - \phi_x^*(t)\}$) and the constants C_x being all equal to 2.*

A simple setting where the conditions of Proposition 2.1 are satisfied is when the functions $\phi_N(x,\cdot)$, $N = 1, 2, \ldots$, are bounded by the log-moment generating function of some random variable W_x (i.e., $\phi_x^*(t) = \log \mathbb{E}[e^{tW_x}]$) such that $\mathbb{E}[W_x] = g(x)$. Clearly, condition (i) holds in that case. Moreover, if there exists a neighborhood $\mathcal{N}$ of zero such that $\phi_x^*(\cdot)$ is finite on $\mathcal{N}$, then it is well known that ϕ_x^* is infinitely differentiable on $\mathcal{N}$ and (iii) holds. In that case, Proposition 1 in Shapiro, Homem-de-Mello, and Kim (2002) ensures that ϕ_x^* is strictly convex on $\mathcal{N}$.

Note that when the samples $\{\boldsymbol{\xi}^i\}$ are i.i.d. we have

$$\phi_N(x,t) \;=\; \frac{1}{N}\log(\mathbb{E}[e^{tN\hat{g}_N(x)}]) \;=\; \frac{1}{N}\log(\{\mathbb{E}[e^{tG(x,\,)}]\}^N) \;=\; \log(\mathbb{E}[e^{tG(x,\,)}]) \;=\; \log M_x(t),$$

where $M_x(t) := \mathbb{E}[e^{tG(x,\,)}]$ is the moment generating function of $G(x,\boldsymbol{\xi})$ evaluated at t. In that case, of course, we have $\phi_N(x,t) = \phi_x^*(t)$ for all N, and the resulting function I_x in Proposition 2.1 is the rate function associated with $G(x,\boldsymbol{\xi})$. Inequality (2.2) then yields the well-known Chernoff upper bounds on the deviation probabilities. It is also well-known

(Cramér's Theorem) that in that case $\gamma_x(\delta)$ is an *asymptotically exact* rate, in the sense that (2.3) holds with equality.

One important consequence of the above developments is the following: Suppose that the function ϕ_x^* in Proposition 2.1 is dominated by the log-moment generating function of the random variable $G(x, \boldsymbol{\xi})$, i.e., $\phi_x^*(t) \leq \phi_x^{\mathrm{MC}}(t) := \log \mathbb{E}[e^{tG(x,\,)}]$. This immediately implies that the rate function I_x dominates the rate function associated with the random variable $G(x, \boldsymbol{\xi})$, which as seen earlier is the asymptotically exact rate function obtained with i.i.d. (i.e., Monte Carlo) sampling. In other words, if one uses a sampling technique that yields functions $\phi_N(x, \cdot)$ for which one can find ϕ_x^* in Proposition 2.1 such that $\phi_x^*(\cdot) \leq \phi_x^{\mathrm{MC}}(\cdot)$, then *the pointwise convergence rate for this sampling technique — in the sense of* (2.2) *— is at least as good as the rate obtained with standard Monte Carlo.* We will use this basic argument repeatedly in the course of this paper.

Under the above conditions, we have the following result. Recall that $\hat{x}_N$ is an optimal solution of (1.3) and S^* is the set of optimal solutions of (1.1). Below, $\mathrm{dist}(z, A)$ denotes the usual Euclidean distance function between a point z and a set A, i.e., $\mathrm{dist}(z, A) := \inf_{y \in A} \|z - y\|$.

Theorem 2.1. *Consider problem* (1.3)*, and suppose that Assumptions A1 and B1 hold.*

1. *Suppose that Assumption A2 holds, and that the random variable $L(\boldsymbol{\xi})$ in Assumption A2 satisfies the large deviations condition in Assumption B1 with $N^{-1}\sum_{j=1}^{N} L(\boldsymbol{\xi}^j)$ and $\mathbb{E}[L(\boldsymbol{\xi})]$ in the role of respectively $\hat{g}_N(x)$ and $g(x)$.*

   Then, given $\varepsilon > 0$, there exist constants $K > 0$ and $\alpha > 0$ such that

$$P\left(\mathrm{dist}(\hat{x}_N, S^*) \geq \varepsilon\right) \;\leq\; Ke^{-\alpha N} \qquad \text{for all } N \geq 1.$$

2. *Suppose that Assumption A3 holds. Then, there exist constants $K > 0$ and $\alpha > 0$ such that*

$$P\left(\hat{x}_N \notin S^*\right) \;\leq\; Ke^{-\alpha N} \qquad \text{for all } N \geq 1.$$

In either case, the constants K and α depend on the random sample used to generate $\hat{g}_N(\cdot)$ only through respectively the constants C_x and the exponent functions $\gamma_x(\cdot)$ in (2.2)*.*

The proof of Theorem 2.1 will be based on the following lemma:

Lemma 2.1. *Suppose that Assumption B1 holds, and that either (i) the set X is finite, or (ii) the conditions in case 1 of Theorem 2.1 hold. Then, for any $\delta > 0$ there exist positive constants $A = A(\delta)$ and $\alpha = \alpha(\delta)$ such that*

$$P\left(|\hat{g}_N(x) - g(x)| \geq \delta\right) \;\leq\; Ae^{-\alpha N}, \qquad \text{for all } x \in X \text{ and all } N \geq 1. \tag{2.5}$$

Moreover, there exists a positive constant K (also dependent on δ) such that

$$P\left(|\hat{g}_N(x) - g(x)| < \delta \text{ for all } x \in X\right) \;\geq\; 1 - Ke^{-\alpha N} \qquad \text{for all } N \geq 1. \tag{2.6}$$

Proof. When X is finite, we can set $\alpha := \inf_{x\in X} \gamma_x(\delta)$ in (2.2) to show (2.5) (with $A := \sup_{x\in X} C_x$). But when X is infinite, in principle we cannot guarantee that such quantity will be strictly positive, so we need a different argument. Let $\eta := \delta/(3\mathbb{E}[L(\boldsymbol{\xi})] + \delta)$, and denote by $B(x,\eta)$ the open ball with center x and radius η. Let $\mathcal{X} = \{x_1, \ldots, x_r\}$ be a collection of points in X such that $X \subset \cup_{k=1}^{r} B(x_k, \eta)$. Notice that the existence of $\mathcal{X}$ is ensured by the compactness of X.

Consider now an arbitrary point $x \in X$. By construction, there exists some $x_k \in \mathcal{X}$ such that $\|x - x_k\| < \eta$. Thus, from (2.1) we have that

$$|\hat{g}_N(x) - \hat{g}_N(x_k)| \;\leq\; \frac{1}{N}\sum_{j=1}^{N} \left|G(x,\boldsymbol{\xi}^j) - G(x_k,\boldsymbol{\xi}^j)\right| \;<\; \hat{L}_N\eta \;=\; \frac{\delta}{3}\,\frac{\hat{L}_N}{\mathbb{E}[Ł(\boldsymbol{\xi})] + \delta/3} \tag{2.7}$$

$$|g(x) - g(x_k)| \;\leq\; \mathbb{E}\left[|G(x,\boldsymbol{\xi}) - G(x_k,\boldsymbol{\xi})|\right] \;<\; \mathbb{E}[Ł(\boldsymbol{\xi})]\eta \;<\; \delta/3. \tag{2.8}$$

Moreover, by Assumption B1 (applied to both $\hat{g}_N(x_k)$ and $\hat{L}_N$) we have that

$$P\left(|\hat{g}_N(x_k) - g(x_k)| \geq \delta/3\right) \;\leq\; C_{x_k}\, e^{-N\gamma_{x_k}(\delta/3)} \tag{2.9}$$

$$P(|\hat{L}_N - \mathbb{E}[L(\boldsymbol{\xi})]| \geq \delta/3) \;\leq\; C_L\, e^{-N\gamma_L(\delta/3)}, \tag{2.10}$$

where C_L and $\gamma_L(\cdot)$ are the quantities given by Assumption B1 when applied to $\hat{L}_N$. Finally, since

$$|\hat{g}_N(x) - g(x)| \;\leq\; |\hat{g}_N(x) - \hat{g}_N(x_k)| + |\hat{g}_N(x_k) - g(x_k)| + |g(x) - g(x_k)|,$$

it follows that

$$\begin{aligned}\{|\hat{g}_N(x) - g(x)| < \delta\} \;&\supseteq\; \{|\hat{g}_N(x) - \hat{g}_N(x_k)| < \delta/3\} \cap \{|\hat{g}_N(x_k) - g(x_k)| < \delta/3\} \\ &\quad \cap \{|g(x_k) - g(x)| < \delta/3\} \\ &\supseteq\; \{|\hat{L}_N - \mathbb{E}[L(\boldsymbol{\xi})]| < \delta/3\} \cap \{|\hat{g}_N(x_k) - g(x_k)| < \delta/3\}\end{aligned} \tag{2.11}$$

and then from (2.9)-(2.10) we have that

$$\begin{aligned}P\left(|\hat{g}_N(x) - g(x)| \geq \delta\right) \;&\leq\; P\left(|\hat{g}_N(x_k) - g(x_k)| \geq \delta/3\right) + P(|\hat{L}_N - \mathbb{E}[L(\boldsymbol{\xi})]| \geq \delta/3) \\ &\leq\; C_{x_k}\, e^{-N\gamma_{x_k}(\delta/3)} + C_L\, e^{-N\gamma_L(\delta/3)}.\end{aligned}$$

By taking $\alpha := \min\left(\min_{k=1,\ldots,r}\{\gamma_{x_k}(\delta/3)\},\ \gamma_L(\delta/3)\right)$ and $A := 2\max\left(\max_{k=1,\ldots,r}\{C_{x_k}\},\ C_L\right)$, inequality (2.5) follows.

To show (2.6), notice that from (2.11) we have

$$\begin{aligned}&P(|\hat{g}_N(x) - g(x)| < \delta \text{ for all } x \in X) \\ &\qquad\geq\; P\left(\{|\hat{g}_N(x_k) - g(x_k)| < \delta/3, \quad k = 1,\ldots,r\} \cap \{|\hat{L}_N - \mathbb{E}[L(\boldsymbol{\xi})]| < \delta/3\}\right) \\ &\qquad\geq\; 1 - \sum_{k=1}^{r} P\left(|\hat{g}_N(x_k) - g(x_k)| \geq \delta/3\right) - P(|\hat{L}_N - \mathbb{E}[L(\boldsymbol{\xi})]| \geq \delta/3),\end{aligned} \tag{2.12}$$

where the last inequality stems from a direct application of Bonferroni's inequality. It follows from (2.9), (2.10) and (2.12) that

$$P\left(|\hat{g}_N(x) - g(x)| < \delta \text{ for all } x \in X\right) \;\geq\; 1 - \frac{r+1}{2} Ae^{-\alpha N}.$$

The proof of (2.6) when X is finite follows a very similar argument and is therefore omitted. □

We return now to the proof of Theorem 2.1.

Proof. Consider first the setting of case 1 of the theorem. Let $\varepsilon > 0$ be given. As mentioned earlier, Assumption A2 implies the existence of some $\delta > 0$ such that $\text{dist}(\hat{x}_N, S^*) < \varepsilon$ whenever $|\hat{g}_N(x) - g(x)| < \delta$ for all $x \in X$; see, e.g., (Rubinstein and Shapiro, 1993, p. 69) for a proof.

Next, suppose that X is finite. Let δ be defined as $(1/2)\min_{x \in X \setminus S^*} g(x) - \nu^*$. By Assumption A1, it is clear that, if $|\hat{g}_N(x) - g(x)| < \delta$ for all $x \in X$, we have that $\hat{g}_N(x) < \hat{g}_N(y)$ for all $x \in S^*$ and all $y \in X \setminus S^*$, i.e., $\hat{x}_N \in S^*$. Now suppose that the conditions in part (ii) of Assumption A3 hold. Then, from Lemma 2.4 in Shapiro and Homem-de-Mello (2000) we know that there exists a finite set of points $\{x_1, \ldots, x_\ell\} \cup \{y_1, \ldots, y_q\}$ such that $x_i \in S^*$, $y_j \in X \setminus S^*$ and, if $\hat{g}_N(x_i) < \hat{g}_N(y_j)$ for all $i \in \{1, \ldots, \ell\}$ and all $j \in \{1, \ldots, q\}$, then $\hat{x}_N \in S^*$ (in fact, the set S_N forms a face of S^*). Therefore, we can use the same argument as in the case where X is finite. We remark that similar results were derived in Kleywegt et al. (2001) and Shapiro and Homem-de-Mello (2000) in the i.i.d. context.

In either case, by Lemma 2.1 the event $\{|\hat{g}_N(x) - g(x)| < \delta \text{ for all } x \in X\}$ occurs with probability at least $1 - Ke^{-\alpha N}$ (where both K and α depend on δ). It follows that in case 1 we have

$$P\left(\text{dist}(\hat{x}_N, S^*) \geq \varepsilon\right) \;\leq\; Ke^{-\alpha N}$$

whereas in case 2 we have

$$P\left(\hat{x}_N \notin S^*\right) \;\leq\; Ke^{-\alpha N},$$

as asserted. Notice that in either case δ does not depend on the particular approximation $\hat{g}_N(\cdot)$; therefore, the constants K and α depend on $\hat{g}_N(\cdot)$ only through respectively the constants C_x and the exponent functions $\gamma_x(\cdot)$ in Assumption B1. □

In essence, Theorem 2.1 says that the existence of an exponential rate of convergence for *pointwise estimators* is enough to ensure an exponential rate of convergence for *optimal solutions* of the corresponding approximating problems, regardless of the sampling scheme adopted. Although reasonably intuitive, such result had not — to the best of our knowledge — been stated or proved anywhere in the literature.

It is important to remark that the last assertion of Theorem 2.1 suggests that a better pointwise convergence rate leads to a better rate of convergence of optimal solutions. Indeed, suppose one has at hand two families of approximations, say, $\{\bar{g}_N(x)\}$ and $\{\tilde{g}_N(x)\}$, whose respective exponent functions $\bar{\gamma}_x(\cdot)$ and $\tilde{\gamma}_x(\cdot)$ in (2.2) are such that $\bar{\gamma}_x(\cdot) \geq \tilde{\gamma}_x(\cdot)$ for all $x \in X$.

Then, the corresponding constants $\bar{\alpha}$ and $\tilde{\alpha}$ will be such that $\bar{\alpha} \geq \tilde{\alpha}$, which *suggests* that the family $\{\bar{g}_N(\cdot)\}$ yields a better rate of convergence of $\hat{x}_N$ to S^*. Of course, Theorem 2.1 only gives an *upper bound* on the deviation probabilities $P(\hat{x}_N \notin S^*)$ and $P(\text{dist}(\hat{x}_N, S^*) \geq \varepsilon)$, so no definitive statements can be made.

Nevertheless, we shall see later specific situations where the pointwise rate of convergence yields an asymptotically exact rate of convergence for the optimization problem; in those cases, superiority of one sampling scheme over another can be established.

An analogous form of Theorem 2.1 can be derived in case Assumption B1′ holds instead of B1. We state the result below for completeness; the proof follows very similar steps to the proof of Theorem 2.1 and is therefore omitted.

Theorem 2.2. *Consider problem* (1.3)*, and suppose that Assumptions A1 and B1′ hold.*

1. *Suppose that Assumption A2 holds, and that the random variable $L(\boldsymbol{\xi})$ in Assumption A2 satisfies the large deviations condition in Assumption B1′ with $N^{-1}\sum_{j=1}^{N} L(\boldsymbol{\xi}^j)$ and $\mathbb{E}[L(\boldsymbol{\xi})]$ in the role of respectively $\hat{g}_N(x)$ and $g(x)$.*

   Then, given $\varepsilon > 0$, there exist a constant $\alpha > 0$ such that

$$\limsup_{N\to\infty} \frac{1}{N} \log P(\text{dist}(\hat{x}_N, S^*) \geq \varepsilon) \;\leq\; -\alpha. \tag{2.13}$$

2. *Suppose that Assumption A3 holds. Then, there exists a constant $\alpha > 0$ such that*

$$\limsup_{N\to\infty} \frac{1}{N} \log P(\hat{x}_N \notin S^*) \;\leq\; -\alpha. \tag{2.14}$$

2.2 Convergence of approximating values

We consider now the convergence of the optimal value of (1.3). In the previous section we showed that a exponential rate of convergence for pointwise estimators leads to an exponential rete of convergence for solutions of (1.3); here, we will show that, in the context of Assumption A3, a Central Limit Theorem-type result for pointwise estimators leads to a Central Limit Theorem-type result for the optimal value of (1.3). Outside of the context of A3, however, one needs more than CLT for pointwise estimators.

We start by making the following probabilistic assumptions on the estimators $\{\hat{g}_N(x)\}$:

Assumption B2: For each $x \in S^*$, the random variable $W_N(x)$ defined as

$$W_N(x) \;:=\; \frac{\hat{g}_N(x) - g(x)}{\sigma_N(x)}, \tag{2.15}$$

where $\sigma_N^2(x) := \text{Var}[\hat{g}_N(x)]$, is such that $W_N(x)$ converges in distribution to a standard Normal (denoted $W_N(x) \xrightarrow{d} \text{Normal}(0,1)$).

Of course, Assumption B2 holds in case of i.i.d. sampling under very mild assumptions — in that case it corresponds to the classical Central Limit Theorem (with $\sigma_N(x) = \sqrt{\text{Var}[G(x,\boldsymbol{\xi})]/N}$). However, as we shall see later, B2 holds in other contexts as well. Note that we impose Assumption B2 only on the set S^* of optimal solutions to (1.1).

The lemma below states a property that will be used in the sequel. In the lemma, the statement "w.p.1 for N large enough" means that, with probability one, there exists an N_0 such that, on each sample path of the underlying process, the condition holds for all $N > N_0$. The value of such N_0 depends on the particular sample path.

Lemma 2.2. *Suppose Assumptions A1 and A3 hold. Then,*

$$\hat{g}_N(\hat{x}_N) - \min_{x^* \in S^*} \hat{g}_N(x^*) \;=\; 0 \quad \textit{w.p.1 for } N \textit{ large enough.}$$

Proof. We have already seen in the proof of Theorem 2.1 that, under Assumptions A1 and A3, we have that $\hat{x}_N \in S^*$ w.p.1 for N large enough. Consider now an arbitrary sample path where such a condition holds. Then, there exists N_0 such that $\hat{x}_N \in S^*$ for all $N > N_0$. That is, for each $N > N_0$ there exists some point $x^*(N) \in S^*$ such that $\hat{x}_N = x^*(N)$. It follows that

$$\hat{g}_N(\hat{x}_N) - \hat{g}_N(x^*(N)) = 0 \quad \text{for all } N > N_0.$$

By definition, $\hat{x}_N$ minimizes $\hat{g}_N(\cdot)$ over X. Together with above equality, this implies that

$$\hat{g}_N(\hat{x}_N) \;\leq\; \min_{x^* \in S^*} \hat{g}_N(x^*) \;\leq\; \hat{g}_N(x^*(N)) \;=\; \hat{g}_N(\hat{x}_N) \quad \text{for all } N > N_0$$

and hence

$$\hat{g}_N(\hat{x}_N) - \min_{x^* \in S^*} \hat{g}_N(x^*) \;=\; 0 \quad \text{for all } N > N_0. \quad \square$$

We then have the following result for rates of convergence:

Theorem 2.3. *Consider problem* (1.3)*, and suppose that Assumptions A1 and A3 hold. Suppose also that the estimators $\hat{g}_N(x)$ have the* same variance *on the set S^* of optimal solutions to* (1.1)*, i.e., the function $\sigma_N^2(\cdot)$ is constant on S^*, and let $(\sigma_N^*)^2$ denote that common value. Then,*

$$\frac{\hat{\nu}_N - \nu^*}{\sigma_N^*} - \min_{x^* \in S^*} W_N(x^*) \;\xrightarrow{d}\; 0. \tag{2.16}$$

If, in addition, Assumption B2 holds and problem (1.1) *has a* unique *optimal solution (call it x^*), then*

$$\frac{\hat{\nu}_N - \nu^*}{\sigma_N(x^*)} \;\xrightarrow{d}\; \textit{Normal}(0,1). \tag{2.17}$$

Proof. By Lemma 2.2 we have that

$$\frac{\hat{g}_N(\hat{x}_N) - \nu^*}{\sigma_N^*} - \frac{\min_{x^* \in S^*} \hat{g}_N(x^*) - \nu^*}{\sigma_N^*} \;=\; 0 \quad \text{w.p.1 for } N \text{ large enough.}$$

Since convergence w.p.1 implies convergence in distribution, it follows that

$$\frac{\hat{g}_N(\hat{x}_N) - \nu^*}{\sigma_N^*} - \frac{\min_{x^* \in S^*} \hat{g}_N(x^*) - \nu^*}{\sigma_N^*} \;\xrightarrow{d}\; 0,$$

and hence

$$\frac{\hat{g}_N(\hat{x}_N) - \nu^*}{\sigma_N^*} - \min_{x^* \in S^*} \frac{\hat{g}_N(x^*) - \nu^*}{\sigma_N^*} \xrightarrow{d} 0.$$

Note that the term inside the min operation is actually $W_N(x^*)$. Moreover, by definition $\hat{g}_N(\hat{x}_N) = \hat{\nu}_N$, which then shows (2.16).

Suppose now that B2 holds and that $S^* = \{x^*\}$. Then, since $W_N(x^*) \xrightarrow{d} \text{Normal}(0,1)$, using a classical result in convergence of distributions (see, e.g., Billingsley 1995, Theorem 25.4) we conclude that

$$\frac{\hat{\nu}_N - \nu^*}{\sigma_N(x^*)} \xrightarrow{d} \text{Normal}(0,1). \quad \square$$

The above result can be slightly strengthened in case the set S^* is finite (say, $S^* = \{x^1, \ldots, x^\ell\}$) and a multivariate version of Assumption B2 holds — namely, that for some deterministic sequence $\{\tau_N\}$ such that $\tau_N \to \infty$ the multivariate process $\tau_N(\hat{g}_N(x^1) - g(x^1), \ldots, \hat{g}_N(x^\ell) - g(x^\ell))$ converges in distribution to a random vector Y with Normal distribution with mean vector zero and covariance matrix Σ. In that case, using a very similar argument to that used by Kleywegt et al. (2001), one can show directly that $\tau_N(\hat{\nu}_N - \nu^*)$ converges in distribution to $\min_{x^* \in S^*} Y(x^*)$. We chose to present our result in the above form because it only requires a univariate CLT.

As mentioned earlier, outside the context of Assumption A3 stronger conditions are required. One possibility is to assume that Assumption A2 holds and that a version of Assumption B2 for functional spaces holds for the space $C(X)$ of continuous functions defined on X. As discussed in Shapiro (1991), Assumption A2 suffices to ensure that each $G(\cdot, \boldsymbol{\xi})$ is a random element of the space $C(X)$, and hence $\hat{g}_N(\cdot) := N^{-1} \sum_{j=1}^N G(\cdot, \boldsymbol{\xi})$ is also a random element of $C(X)$. The validity of a CLT in that functional space, in turn, implies that a convergence result such as (2.17) holds. This approach works well in the i.i.d. context, see Shapiro (1991) for a discussion. However, we are not aware of other contexts where a CLT in a functional space exists, so we do not elaborate further on this topic.

3 Applications

3.1 Latin Hypercube Sampling

Stratified sampling techniques have been used in Statistics and Simulation for years (see Bratley et al. (1987) and Fishman (1997) for references). Generally speaking, the idea is to partition the sample space and fix the number of samples on each partition, which should be proportional to the probability of the partition. This way we ensure that the number of sampled points on each region will be approximately equal to the *expected* number of points to fall in that region. It is intuitive that such procedure yields better variance than crude Monte Carlo; for proofs, see e.g. Fishman (1997). Notice however that, though theoretically appealing, implementing such procedure is far from trivial, since the difficulty is to determine the partitions as well as to compute the corresponding probabilities.

There are many variants of this basic method, one of the most well-known being the so-called *Latin Hypercube Sampling* (LHS), introduced by McKay et al. (1979). The LHS method operates as follows: suppose we want to draw N samples from a random vector $\boldsymbol{\xi}$ with s independent components $\xi_1, \ldots, \xi_s$, each of which has a Uniform(0,1) distribution. The algorithm consists repeating the two steps below for each dimension $j = 1, \ldots, s$:

1. Generate
$$Y^1 \sim U\left(0, \frac{1}{N}\right),\ Y^2 \sim U\left(\frac{1}{N}, \frac{2}{N}\right), \ldots, Y^N \sim U\left(\frac{N-1}{N}, 1\right);$$

2. Let $\xi_j^i := Y^{\pi(i)}$, where π is a random permutation of $1, \ldots, N$.

McKay et al. (1979) show that each sample ξ_j^i (viewed as a random variable) has *the same distribution* as ξ_j, which in turn implies the estimators generated by the LHS method are unbiased. In case of arbitrary distributions, the above procedure is easily modified by drawing the sample as before and applying the *inversion method* to generate the desired random variates.

McKay et al. (1979) also show that, under some conditions, the LHS method does indeed reduce the variance compared to crude Monte Carlo. Stein (1987) shows that, asymptotically (i.e. as the sample size N goes to infinity), LHS is never worse than crude Monte Carlo, even without the assumptions of McKay et al. (1979). More specifically, Owen (1998) shows that $V_{LHS} \leq N/(N-1)V_{MC}$, where V_{LHS} and V_{MC} are respectively the variances under LHS and crude Monte Carlo.

3.1.1 Exponential rate of convergence

In what follows we assume that the components of the random vector $\boldsymbol{\xi}$ are independent. Suppose now the objective function $g(\cdot)$ in (1.1) is approximated by a sample average calculated using the LHS method, i.e., for each $i = 1, \ldots, s$, $\xi_i^1, \ldots, \xi_i^N$ are samples of ξ_i (the ith component of $\boldsymbol{\xi}$) constructed using the LHS method. Call the resulting estimator in (1.2) $\hat{g}_N^{\mathrm{LHS}}(x)$. To study convergence properties of the approximating problem in (1.3), we shall use the tools of section 2. Our goal is to show that the family $\{\hat{g}_N^{\mathrm{LHS}}(\cdot)\}$ satisfies assumption B1, so that we can apply Theorem 2.1 to ensure exponential rate of convergence.

We shall restrict our attention to functions satisfying the following assumption.

Assumption C1: For each $x \in X$, the function $G(x, \cdot)$ is monotone in each component. That is, for each $i = 1, \ldots, s$ and each $\delta > 0$ we have

$$\text{either} \qquad G(x, z + \delta e_i) \geq G(x, z) \qquad \text{for all } z \in \mathbb{R}^s \tag{3.1}$$
$$\text{or} \qquad G(x, z + \delta e_i) \leq G(x, z) \qquad \text{for all } z \in \mathbb{R}^s, \tag{3.2}$$

where as customary e_i denotes the vector with 1 in the ith component and zeros otherwise.

An important case where such an assumption is satisfied is that of two-stage stochastic linear programs with fixed recourse. In section 4 we discuss that case in detail.

An alternative (but stronger) assumption is the following:

Assumption C1′: For each $x \in X$, the function $G(x,\cdot)$ is separable in its components, i.e., there exist functions $G_1, \ldots, G_s$ (all of them mapping $\mathbb{R}^n \times \mathbb{R}$ to $\mathbb{R}$) such that $G(x,\boldsymbol{\xi}) = G_1(x,\xi_1) + \ldots + G_s(x,\xi_s)$. Moreover, $|\mathbb{E}[G_j(x,\xi_j)]| < \infty$, $G(x,\cdot)$ has at most a finite number of singularities, and the set of points at which $G(x,\cdot)$ is discontinuous has Lebesgue measure zero.

The importance of Assumptions C1 and C1′ in the present context is given by the results below:

Theorem 3.1. *Suppose that* (i) *Assumption C1 holds, and* (ii) *for each $x \in X$, the moment generating function of $G(x,\boldsymbol{\xi})$ (denoted $\phi_x^{MC}(t) := \mathbb{E}[e^{tG(x,\,)}]$) is finite everywhere. Consider the LHS estimators $\hat{g}_N^{LHS}(\cdot)$ above defined and the corresponding problem $\min_{x\in X} \hat{g}_N^{LHS}(x)$. Let $\hat{x}_N^{LHS}$ denote an optimal solution of that problem. Then,*

1. *If Assumption A2 holds, then given $\varepsilon > 0$ there exists constants $\tilde{K} > 0$ and $\tilde{\alpha} > 0$ such that*

$$P\left(\operatorname{dist}(\hat{x}_N^{LHS}, S^*) \geq \varepsilon\right) \;\leq\; \tilde{K}e^{-\tilde{\alpha}N} \qquad \text{for all } N \geq 1. \tag{3.3}$$

2. *If Assumption A3 holds, then there exist a constant $\tilde{\alpha} > 0$ such that*

$$P\left(\hat{x}_N^{LHS} \notin S^*\right) \;\leq\; \tilde{K}e^{-\tilde{\alpha}N} \qquad \text{for all } N \geq 1. \tag{3.4}$$

Moreover, in either case the exponent $\tilde{\alpha}$ is at least as large as the corresponding exponent obtained for standard Monte Carlo.

Proof. Note initially that condition (ii) implies that $\mathbb{E}[G(x,\boldsymbol{\xi})^2] < \infty$ and hence Assumption A1 holds under LHS (see Loh 1996). Let $\phi_N(x,t) := \frac{1}{N}\log \mathbb{E}\left[e^{tN\hat{g}_N^{\mathrm{LHS}}(x)}\right]$. If conditions (i) and (ii) above hold, then by Proposition 6 in Drew and Homem-de-Mello (2005) we have that $\phi_N(x,t) \leq \phi_x^{\mathrm{MC}}(t)$ for all x and all t and hence it follows from Proposition 2.1 that Assumption B1 holds for $\{\hat{g}_N^{\mathrm{LHS}}(\cdot)\}$. The two cases of the theorem then parallel the two cases of Theorem 2.1, which shows (3.3) and (3.4).

The last assertion of the theorem is a consequence of the remark following the proof of Theorem 2.1. Indeed, the arguments in the previous paragraph show that the constants C_x and the exponent functions $\gamma_x(\cdot)$ in (2.2) are the same for both LHS and standard Monte Carlo. □

Although Theorem 3.1 only guarantees the same bounds for both LHS and standard Monte Carlo, a closer look at the proof of the inequality $\phi_N(x,\cdot) \leq \phi_x^{\mathrm{MC}}(\cdot)$ shows that this inequality tends to be strict and hence LHS tends to behave better than Monte Carlo.

In case Assumption C1′ holds instead of C1, we have the following stronger result:

Theorem 3.2. *Suppose that the assumptions of Theorem 3.1 are satisfied, but Assumption C1′ holds instead of C1. Then, the conclusions of Theorem 3.1 hold. In addition, we have*

1. *If Assumption A2 holds, then*

$$\lim_{N\to\infty} \frac{1}{N} \log P\left(\operatorname{dist}(\hat{x}_N^{LHS}, S^*) \geq \varepsilon\right) \;=\; -\infty. \tag{3.5}$$

2. If Assumption A3 holds, then

$$\lim_{N\to\infty} \frac{1}{N} \log P\left(\hat{x}_N^{LHS} \notin S^*\right) \;=\; -\infty. \tag{3.6}$$

Proof. The proof of the first part of the theorem follows the same steps as the proof of Theorem 3.1 (except that Proposition 4 in Drew and Homem-de-Mello (2005) is invoked instead of Proposition 6).

To show the second part, consider the inverse of the cumulative distribution function F_j of ξ_j, defined as $F_j^{-1}(u) := \inf\{y \in \Xi_j \,:\, F_j(y) \geq u\}$, where Ξ_j denotes the support of the distribution F_j. Then, by writing each random variable ξ_j as $F_j^{-1}(U_j)$ (where $U_j \sim U(0,1)$), we have that conditions (i) and (ii) of Theorem 3.1 ensure that the assumptions of Theorem 2 in Drew and Homem-de-Mello (2005) are satisfied. The latter result, in turn, ensures that Assumption B1′ holds with the function $\gamma_x = \infty$ everywhere except at zero, where it is equal to zero. Then, (3.5) and (3.6) follow from (2.13) and (2.14) in Theorem 2.2. □

The strength of Theorem 3.2, of course, lies in the asymptotic results (3.5)-(3.6), which show that in the separable case the rate of convergence under LHS is *superexponential*.

3.1.2 Central Limit Theorem

We study now the convergence of optimal values of the approximating problem (1.3) under LHS. To do so we shall apply results of section 2.2. Before that, however, we need to review some results related to the ANOVA decomposition of a function.

Let $U = (U_1, \ldots, U_s)$ be an s-dimensional random vector with uniform distribution on $[0,1]^s$, $f : [0,1]^s \longrightarrow \mathbb{R}$ an arbitrary function and consider the problem of estimating $I := \mathbb{E}[f(U)]$. Stein (1987) shows that, when $\mathbb{E}[(f(U))^2] < \infty$, f can be decomposed as

$$f(U) \;=\; \mathbb{E}[f(U)] + \sum_{k=1}^{s} f_k(U_k) + r(U), \tag{3.7}$$

where $f_k(U_k) = \mathbb{E}[f(U) \,|\, U_k] - \mathbb{E}[f(U)]$ and $r(U)$ is the *residual* term, which satisfies $\mathbb{E}[r(U) \,|\, U_j] = 0$ for all j. Moreover, Stein (1987) also shows that the variance of the estimator I_{LHS} (defined as $I_{\text{LHS}} := N^{-1} \sum_{i=1}^{N} f(U^i)$, where $U^1, \ldots, U^N$ are samples drawn with LHS) satisfies

$$\sigma_N^2 := \text{Var}[I_{\text{LHS}}] \;=\; N^{-1}\mathbb{E}\left[(r(U))^2\right] + o(N^{-1}) \tag{3.8}$$

as N goes to infinity.

Using the results in (3.7) and (3.8), Owen (1992) shows that, when f is bounded, a Central Limit Theorem holds for the estimator I_{LHS} under Latin Hypercube Sampling. More specifically, he shows that

$$N^{1/2}(I_{\text{LHS}} - I) \;\xrightarrow{d}\; \text{Normal}(0, \sigma^2), \quad \text{where } \sigma^2 := \mathbb{E}[(r(U))^2]. \tag{3.9}$$

Next, notice that from (3.8) we can write

$$\frac{I_{\text{LHS}} - I}{\sigma_N} \;=\; \frac{N^{1/2}(I_{\text{LHS}} - I)}{\left[\sigma^2 + \frac{o(N^{-1})}{N^{-1}}\right]^{1/2}}.$$

Since $N^{1/2}(I_{\text{LHS}} - I) \xrightarrow{d} \text{Normal}(0, \sigma^2)$ and the deterministic sequence $\{[\sigma^2 + \frac{o(N^{-1})}{N^{-1}}]^{1/2}\}$ converges to σ, it follows from a classical result in probability theory (see, e.g., Chung 1974, p. 93) that, when $\sigma > 0$,

$$\frac{I_{\text{LHS}} - I}{\sigma_N} \xrightarrow{d} \frac{1}{\sigma}\text{Normal}(0, \sigma^2) = \text{Normal}(0, 1). \tag{3.10}$$

Notice that the condition $\mathbb{E}[(f(U))^2] < \infty$ also implies that a Strong Law of Large Numbers holds for LHS, i.e.,

$$|I_{\text{LHS}} - I| \to 0 \quad w.p.1; \tag{3.11}$$

for a proof, see Loh (1996).

By applying (3.10) to our setting we see that Assumption B2 holds for LHS under some boundedness condition, provided that the ANOVA residual of $G(x, \cdot)$ is positive for all $x \in S^*$. Moreover, the same boundedness condition implies, via (3.11), that Assumption A1 holds. Thus, under additional assumptions we can apply Theorem 2.3. As before, we write $\boldsymbol{\xi} = (F_1^{-1}(U_1), \ldots, F_s^{-1}(U_s))$, where $U_1, \ldots, U_s$ are independent uniform random variables in $[0, 1]$. We summarize the result in the theorem below.

Theorem 3.3. *Consider the LHS estimators $\hat{g}_N^{LHS}(\cdot)$ above defined and the corresponding problem $\min_{x \in X} \hat{g}_N^{LHS}(x)$. Let $\hat{x}_N^{LHS}$ and $\hat{\nu}_N^{LHS}$ denote respectively an optimal solution and the optimal value of that problem.*

1. *If Assumption A2 holds, then* $\text{dist}(\hat{x}_N^{LHS}, S^*) \to 0$ *w.p.1 and $\hat{\nu}_N^{LHS} \to \nu^*$ w.p.1.*
2. *If Assumption A3 holds, then $\hat{x}_N^{LHS} \in S^*$ w.p.1 for N large enough and $\hat{\nu}_N^{LHS} \to \nu^*$ w.p.1. In addition, suppose that for each $x \in X$ the function $G(x, \cdot)$ is* bounded *and that the distribution of $\boldsymbol{\xi}$ has* bounded support*. If problem* (1.1) *has a* unique *optimal solution (call it x^*) and the function $G(x^*, \cdot)$ is not separable in its components, then*

$$\frac{\hat{\nu}_N - \nu^*}{\sigma_N(x^*)} \xrightarrow{d} \textit{Normal}(0, 1),$$

where $\sigma_N^2(x^*)$ is the variance of $\hat{g}_N^{LHS}(x^*)$. Moreover, there exists a positive constant C such that

$$\sigma_N^2(x^*) = N^{-1}C + o(N^{-1}) \tag{3.12}$$

as $N \to \infty$.

Theorem 3.3 shows that the rate of convergence of optimal values under LHS (under the conditions of Assumption A3) is $N^{-1/2}$. Thus, compared to standard Monte Carlo we can see that, although LHS will likely reduce the variance of pointwise estimators, it cannot improve the *rate* of convergence unless the function $G(x^*, \cdot)$ is separable in its components, since in that case the residual in the ANOVA decomposition of $G(x^*, \cdot)$ is equal to zero and so we expect the convergence rate to be much faster. Indeed, recall from Theorem 3.2 that, under the assumptions of that theorem (which include separability), convergence of optimal solutions is superexponential. Note also that, when S^* is finite (but not necessarily a singleton) and $G(x^*, \cdot)$ is not separable in its components for all $x^* \in S^*$, the stronger result discussed in the paragraph following the proof of Theorem 2.3 applies with $\tau_N = N^{1/2}$, since the aforementioned CLT result proved by Owen (1992) is also valid in a multivariate context.

3.2 Randomized Quasi Monte Carlo

For completeness, we provide in this section a brief review of quasi-Monte Carlo (QMC) techniques. We follow mostly Niederreiter (1992), which we refer to for comprehensive treatments of QMC concepts. Let U be an s-dimensional random vector with uniform distribution on $[0,1]^s$, $f : [0,1]^s \longrightarrow \mathbb{R}$ an arbitrary function and consider the problem of estimating $I := \mathbb{E}[f(U)]$.

The basic idea of QMC is to calculate a sample average estimate as in the standard Monte Carlo but, instead of drawing a random sample from the uniform distribution on $[0,1]^s$, a certain set of points $u^1, \ldots, u^N$ on space $[0,1]^s$ is carefully chosen. The deterministic estimate

$$I_{\text{QMC}} := \frac{1}{N} \sum_{i=1}^{N} f(u^i) \tag{3.13}$$

is constructed. A key result is the so-called Koksma-Hlawka inequality which, roughly speaking, states that the quality of the approximation given by I_{QMC} depends on the quality of the chosen points (measured by the difference between the corresponding empirical measure and the uniform distribution, which is quantified by the so-called *star-discrepancy*) as well as on the nature of the function f (measured by its total variation). A great deal of the research on QMC methods aims at determining ways to construct *low-discrepancy sequences*, i.e., sequences of points $u^1, u^2, \ldots$ for which the star-discrepancy is small for all N. A particular type of sequence that has proven valuable is defined in terms of (t, m, s)-nets. We need some definitions before delving into more details, which we do next.

Let $b \geq 2$ be an arbitrary integer, called the base. An *elementary interval in base b* (in dimension s) is a subinterval E of $[0,1]^s$ of the form

$$E \;=\; \prod_{j=1}^{s} \left[\frac{a_j}{b^{d_j}},\; \frac{a_j + 1}{b^{d_j}} \right]$$

for nonnegative integers $\{a_j\}$ and $\{d_j\}$ such that $a_j < b^{d_j}$ for all j. The volume of E is $b^{-\sum_j d_j}$. Next, let t and m be nonnegative integers such that $t \leq m$. A finite sequence of b^m points is a *(t, m, s)-net* in base b if every elementary interval in base b of volume b^{t-m} contains exactly b^t points of the sequence. A sequence of points $u^1, u^2, \ldots$ is a *(t, s)-sequence* in base b if, for all integers $k \geq 0$ and $m > t$, the set of points consisting of the u^n such that $kb^m \leq n < (k+1)b^m$ is a (t, m, s)-net in base b.

The advantage of (t, m, s)-nets becomes clear from a result due to (Niederreiter, 1992, Theorems 4.10 and 4.17), who shows that the error $|I_{\text{QMC}} - I|$ is: (i) of order $(\log N)^{s-1}/N$ when I_{QMC} is computed from a (t, m, s)-net in base b with $m > 0$; (ii) of order $(\log N)^s/N$ when I_{QMC} is computed from the first $N \geq 2$ terms of a (t, s)-sequence in base b. Note that in case (i) N must be equal to b^m, whereas in case (ii) N is arbitrary, which explains the weaker bound. In either case, it is clear that, asymptotically, the error is smaller than $N^{-1/2}$ given by standard Monte Carlo methods.

Despite the advantage of QMC with respect to error rates, the method has two major drawbacks:

(a) The bounds provided by the Koksma-Hlawka inequality involve difficult-to-compute quantities such as the total variation of f, i.e., they yield qualitative (rather than quantitative) results; hence, obtaining an exact estimate of the error may be difficult.

(b) A comparison of the functions $(\log N)^s/N$ and $N^{-1/2}$ shows that even though asymptotically the error from QMC is smaller than the error from standard Monte Carlo, such an advantage does not appear until N is very large, unless s is small.

These difficulties have long been realized by the QMC community, and various remedies have been proposed. A common way to overcome difficulty (a) above is to incorporate some randomness into the choice of QMC points. By doing so, errors can be estimated using standard methods, e.g., via multiple independent replications. This is the main idea of *randomized* QMC methods (RQMC), see Fox (2000) and Owen (2000) for detailed discussions.

One particular technique we are interested in using relies on "scrambling" the decimal digits of each point of a (t, s)-sequence in a proper way. This idea was proposed by Owen (1995), and has gained popularity due to the nice properties of the randomized sequence. We shall use these properties below.

3.2.1 Using QMC in optimization

Consider again the family of estimators defined in (1.2), and assume that N is of the form b^m for some positive integer m. Suppose that $\{\boldsymbol{\xi}^i\}$ is generated by a (t, m, s)-net, and call the resulting family $\{\hat{g}_N^{\mathrm{QMC}}(x)\}$.

Let us fix $x \in X$ for a moment. As discussed above, when $\{\boldsymbol{\xi}^i\}$ is generated by a *standard* (i.e., not scrambled) (t, m, s)-net we have

$$|\hat{g}_N^{\mathrm{QMC}}(x) - g(x)| \;=\; O\left(\frac{(\log_b N)^{s-1}}{N}\right), \tag{3.14}$$

provided that $G(x, \cdot)$ is of finite total variation (in the sense of Hardy and Krause). Clearly, this implies that $\hat{g}_N^{\mathrm{QMC}}(x) \to g(x)$ as $N \to \infty$. Now suppose that $\{\boldsymbol{\xi}^i\}$ is generated by a scrambled (t, m, s)-net, and call the corresponding estimator $\hat{g}_N^{\mathrm{RQMC}}$. Owen (1997a) shows that scrambled (t, m, s)-nets are (t, m, s)-nets with probability one, which then implies that

$$\hat{g}_N^{\mathrm{RQMC}}(x) \to g(x) \quad \text{w.p.1.} \tag{3.15}$$

Moreover, $\hat{g}_N^{\mathrm{RQMC}}(x)$ is an unbiased estimator of $g(x)$, i.e., $\mathbb{E}[\hat{g}_N^{\mathrm{RQMC}}(x)] = g(x)$. Notice that the term "with probability one" above refers to the probability space where the random permutations that are part of the scrambling algorithm lie. We assume that this probability space is the same where the random vectors $\boldsymbol{\xi}$ are defined.

For some of the results that follow we will need the following assumption.

Assumption D1: Suppose the following conditions hold for each $x \in S^*$:

$$\left|\frac{\partial^s}{\partial u_1 \dots \partial u_s} G(x, F^{-1}(u_1, \dots u_s)) - \frac{\partial^s}{\partial u_1 \dots \partial u_s} G(x, F^{-1}(v_1, \dots v_s))\right| \;\le\; B\|u - v\|^\beta \tag{3.16}$$

(for some $B > 0$ and some $\beta \in (0, 1]$), and

$$\int_{[0,1]^s} \left[\frac{\partial^s}{\partial u_1 \dots \partial u_s} G(x, F^{-1}(u_1, \dots u_s)) \right]^2 du \; > \; 0. \tag{3.17}$$

In the above, F^{-1} is a mapping from $\in [0, 1]^s$ into $\mathbb{R}^s$ such that $\boldsymbol{\xi} = F^{-1}(U)$ and U is a random vector uniformly distributed on $[0, 1]^s$.

A few remarks about cases where Assumption D1 is satisfied are now in order. Suppose that the components $\xi_1, \dots, \xi_s$ of $\boldsymbol{\xi}$ are mutually independent. As before, we can write $\boldsymbol{\xi} = (F_1^{-1}(U_1), \dots, F_s^{-1}(U_s))$, where $U_1, \dots, U_s$ are independent uniform random variables in $[0, 1]$ and F_j^{-1} is the inverse cdf of ξ_j. Suppose momentarily that G is infinitely differentiable in the second argument and that each F_j^{-1} is differentiable as well. Then, we have that

$$\frac{\partial}{\partial u_1} G(x, F_1^{-1}(u_1), \dots, F_s^{-1}(u_s)) \;=\; \frac{\partial}{\partial \xi_1} G(x, \xi_1, F_2^{-1}(u_2) \dots, F_s^{-1}(u_s)) \Big|_{\xi_1 = F_1^{-1}(u_1)} \frac{\partial}{\partial u_1} F_1^{-1}(u_1)$$

so by repeating the calculation for the higher-order mixed derivatives we obtain that

$$H(u_1, \dots, u_s) \;:=\; \frac{\partial^s}{\partial u_1 \dots \partial u_s} G(x, F_1^{-1}(u_1), \dots, F_s^{-1}(u_s)) \;= \tag{3.18}$$

$$\;=\; \frac{\partial^s}{\partial \xi_1 \dots \partial \xi_s} G(x, \xi_1, \dots, \xi_s) \Big|_{\substack{\xi_j = F_j^{-1}(u_j) \\ j=1,\dots,s}} \frac{\partial}{\partial u_1} F_1^{-1}(u_1) \dots \frac{\partial}{\partial u_s} F_s^{-1}(u_s). \tag{3.19}$$

It follows that, if the gradient of the function H defined in (3.18) is uniformly bounded for all $u \in [0, 1]^s$, then H is Lipschitz (see, e.g, Bartle 1987, Corollary 40.6), i.e., (3.16) holds. A sufficient condition for uniform boundedness of $\nabla H(u)$ on $[0, 1]^s$ is its continuity on $[0, 1]^s$. Equation (3.19) shows that continuous differentiability of G (up to order $s + 1$) and F_j^{-1}, $j = 1, \dots, s$ (up to second order) on the closed set $[0, 1]^s$ suffice for that. Of course, imposing a continuous differentiability assumption on F_j^{-1} restricts the type of distributions that can be used; we shall return to that issue shortly.

Condition (3.17) essentially says that interactions of order up to s are significant, at least on a set of positive probability. For example, (3.17) does not hold if $G(x, \cdot)$ is linear for $x \in S^*$, since the mixed derivatives of any order bigger than 1 are equal to zero. Situations like that suggest that the *effective dimension* of the problem is less than s (Owen, 1997a) — indeed, in the linear case the effective dimension is 1. In that case, one should apply quasi-Monte Carlo only to the most significant variables, for which mutual interaction is significant.

Applying the above results on randomized QMC to the general context of section 2.2 we obtain the following:

Theorem 3.4. *Consider the RQMC estimators $\hat{g}_N^{RQMC}(\cdot)$ above defined and the corresponding problem $\min_{x \in X} \hat{g}_N^{RQMC}(x)$. Let $\hat{x}_N^{RQMC}$ and $\hat{\nu}_N^{RQMC}$ denote respectively an optimal solution and the optimal value of that problem.*

1. *If Assumption A2 holds, then $\mathrm{dist}(\hat{x}_N^{RQMC}, S^*) \to 0$ w.p.1 and $\hat{\nu}_N^{RQMC} \to \nu^*$ w.p.1.*

2. *If Assumption A3 holds, then* $\hat{x}_N^{RQMC} \in S^*$ *w.p.1 for* N *large enough and* $\hat{\nu}_N^{RQMC} \to \nu^*$ *w.p.1. If, in addition, Assumption D1 holds, problem* (1.1) *has a* unique *optimal solution (call it* x^**) and the samples* $\{\boldsymbol{\xi}^i\}$ *are generated by a scrambled* $(0, m, s)$*-net (i.e.,* $t = 0$*), then*

$$\frac{\hat{\nu}_N^{RQMC} - \nu^*}{\sigma_N(x^*)} \xrightarrow{d} Normal(0, 1),$$

where $\sigma_N^2(x^*)$ *is the variance of* $\hat{g}_N^{RQMC}(x^*)$*. Moreover, when* N *is of the form* b^m*, there exist positive constants* c *and* C *such that*

$$c\,\frac{(\log_b N)^{s-1}}{N^3} \;\le\; \sigma_N^2(x^*) \;\le\; C\,\frac{(\log_b N)^{s-1}}{N^3} \tag{3.20}$$

as $m \to \infty$*.*

Proof. Let us fix $x \in X$. The assertion in case 1 and the first assertion in case 2 follow directly from (3.15) (which implies that Assumption A1 holds) and the remark following Assumption A3.

Consider now the random variable $W(x)$ defined as

$$W(x) \;:=\; \frac{\hat{g}_N^{\mathrm{RQMC}}(x) - g(x)}{\sigma_N(x)}, \tag{3.21}$$

where $\sigma_N^2(x) := \mathrm{Var}[\hat{g}_N^{\mathrm{RQMC}}(x)]$. Here we resort to a key result on scrambled (t, m, s)-nets proved by Loh (2003) — building upon previous results by Owen (1997a,b) — that says that a Central Limit Theorem holds for pointwise estimators constructed with a scrambled $(0, m, s)$-net. Assumption D1 essentially translates the conditions in Loh (2003) into our notation. It follows that, under D1, $W(x)$ converges in distribution to the standard normal for each $x \in S^*$, i.e., Assumption B2 holds and hence the conclusion follows from Theorem 2.3. □

Theorem 3.4 shows the benefits of using randomized quasi-Monte Carlo methods in optimization. Essentially, it says that, in the setting of Assumption A3, the convergence rate of optimal values is of order $[(\log_b N)^{s-1}/N^3]^{1/2}$, which asymptotically is much better than the $N^{-1/2}$ obtained with standard Monte Carlo. This suggests that RQMC methods can be very efficacious for stochastic optimization. Note however that, strictly speaking, the theorem applies only to the case where X is finite, since the assumption of finite support of $\boldsymbol{\xi}$ in the second case of Assumption A3 conflicts with the smoothness condition in Assumption D1. We discuss the smoothness issue in more detail next.

Discussion of smoothness

It must be noted that the assumptions of Theorem 3.4 are important to ensure that the convergence rate $[(\log_b N)^{s-1}/N^3]^{1/2}$ is achieved. In particular, it is important that the smoothness condition required by Assumption D1 hold. Without smoothness, not even pointwise estimation yields such a rate. To illustrate this point, consider the estimation of $\mathbb{E}[(Y_1 Y_2 Y_3)^2]$, where $Y_1,\ Y_2,\ Y_3$ are independent random variables with discrete distribution

$P(Y_i = 1) = 1/2$, $P(Y_i = 2) = 1/3$, $P(Y_i = 3) = 1/6$. Clearly, F_i^{-1} is discontinuous. Figure 1 shows the rate of convergence for the pointwise estimators by depicting the logarithm of the standard deviation of the RQMC estimators, computed over 25 independent replications. The figure also shows the rates that are obtained when F_i^{-1} is replaced by a smooth function F_i^{Δ} such that F_i^{-1} and F_i^{Δ} coincide everywhere except on a interval of size 2Δ around each discontinuity point. The figure suggests that the rate of convergence for the original distribution is of order $1/N$, and improves as Δ gets larger, i.e., as F_i^{Δ} gets smoother — indeed, for $\Delta = 0.16$ the rate of convergence essentially follows $[(\log_b N)^{s-1}/N^3]^{1/2}$ (the log of the latter quantity is called "predicted slope" on the graph). Smoothing, however, comes at a price, as the estimators become biased; Figure 2 shows that the bias increases with Δ.

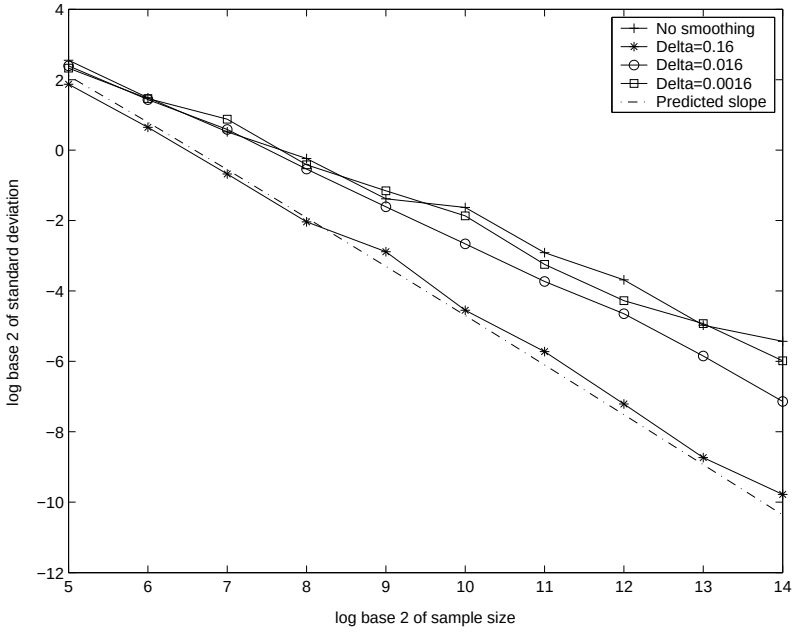

Figure 1: Rates of convergence for pointwise estimation of $\mathbb{E}[(Y_1Y_2Y_3)^2]$.

Despite the above shortcomings, it is important to mention that Theorem 2.3 is valid regardless of any smoothness conditions. As commented earlier, that result essentially says that pointwise rates of convergence carry over to the optimization case. So, if one shows that a Central Limit Theorem holds for RQMC under nonsmooth (or potentially discontinuous) functions — perhaps with a rate of $1/N$ — then Theorem 2.3 will ensure that the optimal value estimators $\hat{\nu}_N^{\mathrm{RQMC}}$ converge at the same rate. The aforementioned result by Loh (2003) is, however, the only CLT-type result available for RQMC, at least to the best of our knowledge.

4 Two-Stage Stochastic Programs

In this section we discuss the application of the results outlined in the previous sections to two-stage stochastic linear programs (see, e.g., Birge and Louveaux 1997 for a comprehensive

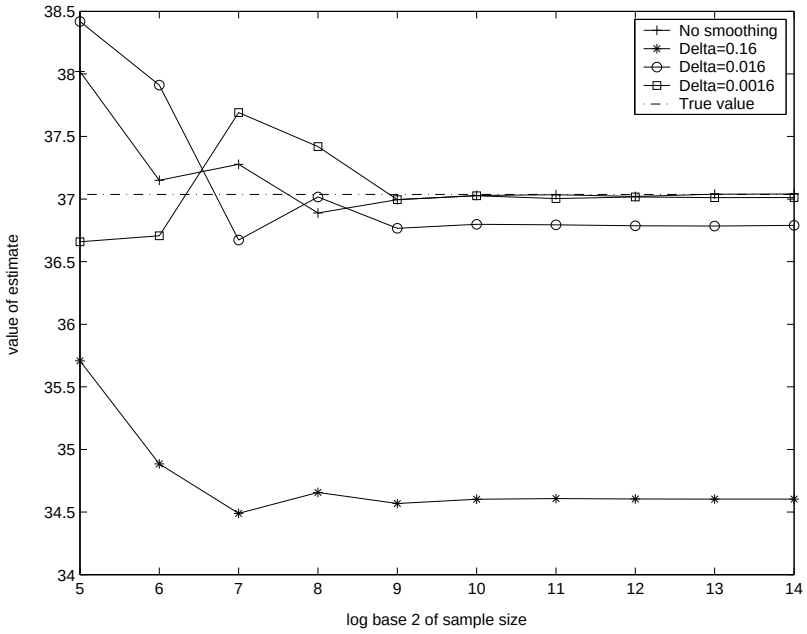

Figure 2: Values of the estimates of $\mathbb{E}[(Y_1Y_2Y_3)^2]$.

discussion of this class of problems). We consider problems of the form

$$\min_{x \in X} c^t x + \mathbb{E}[Q(x, \boldsymbol{\xi})], \tag{4.1}$$

where X is a convex polyhedral set,

$$Q(x, \boldsymbol{\xi}) = \inf \left\{ q^t y \,:\, Wy \leq h - Tx,\ y \geq 0 \right\} \tag{4.2}$$

and $\boldsymbol{\xi} = (h, T)$. As before, $\boldsymbol{\xi}$ is an s-dimensional random vector with arbitrary distribution. Let $G(x, \boldsymbol{\xi})$ denote the function $c^t x + Q(x, \boldsymbol{\xi})$; then, we see that the the above problem falls in the framework of (1.1).

The use of Monte Carlo sampling to solve two-stage problems has been extensively studied in the literature, both from algorithmic (e.g., Higle and Sen 1991, Infanger 1994, Linderoth, Shapiro, and Wright 2005, Shapiro and Homem-de-Mello 1998) and theoretical perspectives (see, for instance, Shapiro 2003 for a compilation of results).

Note that the function $Q(x, \boldsymbol{\xi})$ can be written in the form $Q(x, \boldsymbol{\xi}) = \tilde{Q}(\boldsymbol{\xi} - Tx)$, where

$$\tilde{Q}(z) \;=\; \inf \left\{ q^t y \,:\, Wy \leq z,\ y \geq 0 \right\}. \tag{4.3}$$

By duality, we see that the function $\tilde{Q}(\cdot)$ can be represented in the form

$$\tilde{Q}(z) \;=\; \sup\{u^t z : W^t u \leq q,\ u \geq 0\}. \tag{4.4}$$

For the sake of simplicity we assume that: (i) for every vector z the system $Wy \leq z$, $y \geq 0$, has a solution (the recourse is complete), and (ii) the system $W^t u \leq q,\ u \geq 0$ has a solution (dual feasibility). Under these assumptions, $\tilde{Q}(\cdot)$ is a finite valued, piecewise linear

convex function. This in turn implies that the function $G(x, \boldsymbol{\xi})$ is also piecewise linear convex (simultaneously in both arguments) and can be written as

$$G(x, \boldsymbol{\xi}) \;=\; \max_{k=1,\dots,r} c^t x + (v^k)^t (h - Tx), \tag{4.5}$$

where $v^1, \dots, v^r$ are the vertices of the polyhedron $\{u : W^t u \leq q,\ u \geq 0\}$. Furthermore, by standard subdifferential calculus we have that the subdifferential set of $G(x, \boldsymbol{\xi})$ with respect to x is given by

$$\partial_x G(x, \boldsymbol{\xi}) \;=\; \mathrm{conv}\{c - T^t v^k \,:\, G(x, \boldsymbol{\xi}) = c^t x + (v^k)^t (h - Tx),\ k = 1, \dots, r\}, \tag{4.6}$$

where "conv" denote the convex hull of the set.

In the discussion that follows we assume that the matrix T is deterministic, so that $\boldsymbol{\xi} = h$, and that the feasibility set X is compact.

4.1 LHS results

In order to apply the results for Latin Hypercube Sampling discussed in section 3.1, we need to verify that the corresponding assumptions are satisfied. Consider Assumption A2. It follows from (4.6) that $\partial_x G(x, \boldsymbol{\xi})$ is uniformly bounded for all x and all $\boldsymbol{\xi}$ and thus, by a version of the mean-value theorem for subdifferentiable functions (see, e.g., Hiriart-Urruty and Lemarechal 1993, Theorem VI.2.3.3), we conclude that A2 holds. Next, notice that from (4.3) we have $G(x, \boldsymbol{\xi}) = \min\{q^t y \,:\, Wy \leq \boldsymbol{\xi} - Tx,\ y \geq 0\}$. Thus, for any $\delta > 0$ we have that $G(x, \boldsymbol{\xi} + \delta e_i) \leq\ G(x, \boldsymbol{\xi})$, i.e, Assumption C1 holds.

It follows from the above discussion and from Theorem 3.1 that, if the moment generating function of $G(x, \boldsymbol{\xi})$ is finite everywhere for all x, then given $\varepsilon > 0$ there exist constants $\tilde{K} > 0$ and $\tilde{\alpha} > 0$ such that

$$P\left(\mathrm{dist}(\hat{x}_N^{\mathrm{LHS}}, S^*) \geq \varepsilon\right) \;\leq\; \tilde{K} e^{-\tilde{\alpha} N} \qquad \text{for all } N \geq 1.$$

Moreover, the exponent $\tilde{\alpha}$ is at least as large as the corresponding exponent obtained for standard Monte Carlo. This suggests that convergence under LHS will indeed be faster than under standard Monte Carlo.

As mentioned earlier, $G(\cdot, \boldsymbol{\xi})$ is piecewise linear. Thus, if $\boldsymbol{\xi}$ has finite support then Assumption A3 holds, so from Theorem 3.1 we have that

$$P\left(\hat{x}_N^{\mathrm{LHS}} \notin S^*\right) \;\leq\; \tilde{K} e^{-\tilde{\alpha} N} \qquad \text{for all } N \geq 1.$$

It is fruitful to compare the above result with the i.i.d. case derived in Shapiro and Homem-de-Mello (2000). Indeed, when problem (4.1) has a unique solution x^*, a slightly modified proof of Theorem 3.2 in Shapiro and Homem-de-Mello (2000) shows that there exists $\beta > 0$ such that

$$\limsup_{N \to \infty} \frac{1}{N} \log P\left(\hat{x}_N \neq x^*\right) \;=\; -\beta, \tag{4.7}$$

where $\hat{x}_N$ is the solution obtained with standard Monte Carlo. Moreover, the constant β is given by the minimum of a number of pointwise rates $-\gamma_x(\delta_0)$ in (2.2) (for a fixed $\delta_0 > 0$)

over a *finite* number of x's. Since finite support of $\boldsymbol{\xi}$ implies that the moment generating function of $G(x, \boldsymbol{\xi})$ is finite everywhere for all x, it follows from Proposition 6 in Drew and Homem-de-Mello (2005) that the pointwise rates $-\gamma_x(\delta_0)$ under LHS are no worse than under Monte Carlo. It follows that, when LHS is applied, an equation similar to (4.7) holds and the resulting constant β is *no worse* than under Monte Carlo. Since the rate in (4.7) is exact, we conclude that LHS can only improve upon Monte Carlo in this setting.

Next, we apply Theorem 3.3 to the present context. Clearly, (4.5) implies that $G(x, \cdot)$ is bounded for each x. It follows that, if the distribution of $\boldsymbol{\xi}$ has bounded support, problem (1.1) has a unique optimal solution x^*, and the function $G(x^*, \cdot)$ is not separable in its components, then

$$\frac{\hat{\nu}_N - \nu^*}{\sigma_N(x^*)} \stackrel{d}{\rightarrow} \text{Normal}(0, 1),$$

where $\sigma_N(x^*) := \text{Var}[\hat{g}_N^{\text{LHS}}(x^*)] = N^{-1}C + o(N^{-1})$ for some positive constant C. Note that the non-separability condition is reasonable in this setting, since at the optimal solution x^* typically it happens that the maximum in (4.5) is achieved by more than one k, so $G(x^*, \cdot)$ is not linear.

4.2 QMC results

We now apply the results from section 3.2 to the two-stage stochastic programming model described above. As remarked earlier, Assumption A2 holds in that context. Thus, Theorem 3.4 ensures that $\text{dist}(\hat{x}_N^{\text{RQMC}}, S^*) \rightarrow 0$ w.p.1 and $\hat{\nu}_N^{\text{RQMC}} \rightarrow \nu^*$ w.p.1, a result shown in Koivu (2005) and Pennanen and Koivu (2005). When $\boldsymbol{\xi}$ has finite support (i.e. Assumption A3 holds), we obtain a stronger result, namely, that $\hat{x}_N^{\text{RQMC}} \in S^*$ w.p.1 for N large enough.

The main use of Theorem 3.4, however, is the determination of rates of convergence. Suppose that $\boldsymbol{\xi}$ has finite support. In order to apply that result, we need to show that Assumption D1 holds. Strictly speaking, it is clear that Assumption D1 cannot hold in this case, since $G(x, \cdot)$ is non-differentiable for each x. Moreover, the assumption that $\boldsymbol{\xi}$ has finite support causes the inverse cdfs F_j^{-1} to be discontinuous.

Nevertheless, we can make an "empirical use" of Theorem 3.4 in this context by applying some smoothing techniques, as described in the discussion following Assumption D1. As mentioned there, one can replace the inverse functions F_j^{-1} with smooth approximations F_j^{Δ} such that each F_j^{Δ} is C^2 on $[0, 1]^s$. It remains to show that G can also be approximated by smooth functions. Many smoothing techniques exist; here we describe a particular one, discussed in Guillaume and Seeger (2001), that suits our purposes.

Consider the function $\tilde{Q}(\cdot)$ defined in (4.4). As mentioned earlier, under the assumptions made in this section there exist vectors $v^1, \ldots, v^r$ — the vertices of the dual polyhedron (4.4) — such that $\tilde{Q}(z) = \max_{k=1,\ldots,r} z'v^k$. Let μ denote an *arbitrary* positive discrete probability measure on $\{1, \ldots, r\}$, with $\mu_k := \mu\{k\} > 0$. Let $\mathbf{v}$ be a random variable defined on $v^1, \ldots, v^r$ with law μ, i.e. $P(\mathbf{v} = v^k) = \mu_k$. For each $\tau > 0$ define the function

$$\tilde{Q}_\mu^\tau(z) := \frac{1}{\tau} \log \mathbb{E}_\mu \left[e^{\tau \mathbf{v}' z} \right]. \tag{4.8}$$

As remarked in Guillaume and Seeger (2001), the sequence $\{\tilde{Q}_\mu^\tau\}_\tau$ converges to $\tilde{Q}$ not only pointwise but also uniformly on compact sets. Moreover, $\{\tilde{Q}_\mu^\tau\}_\tau$ is infinitely differentiable, which then allows us to look at the limit of the mixed partial derivatives of $\tilde{Q}_\mu^\tau$.

Let us look initially at the first and second order derivatives of $\tilde{Q}_\mu^\tau$. We have

$$\begin{aligned}
\frac{\partial \tilde{Q}_\mu^\tau}{\partial z_i}(z) &= \frac{\mathbb{E}_\mu\left[\mathbf{v}_i e^{\tau \mathbf{v}' z}\right]}{\mathbb{E}_\mu\left[e^{\tau \mathbf{v}' z}\right]} \\
\frac{\partial^2 \tilde{Q}_\mu^\tau}{\partial z_i \partial z_j}(z) &= \tau \, \frac{\mathbb{E}_\mu\left[\mathbf{v}_i \mathbf{v}_j e^{\tau \mathbf{v}' z}\right] \mathbb{E}_\mu\left[e^{\tau \mathbf{v}' z}\right] - \mathbb{E}_\mu\left[\mathbf{v}_i e^{\tau \mathbf{v}' z}\right] \mathbb{E}_\mu\left[\mathbf{v}_j e^{\tau \mathbf{v}' z}\right]}{\left(\mathbb{E}_\mu\left[e^{\tau \mathbf{v}' z}\right]\right)^2}.
\end{aligned}$$

It is not difficult to calculate higher-order derivatives, though the expressions get rather cumbersome. The key to analyze the behavior of these quantities as $\tau \to \infty$ is to observe that, for large τ, the dominating terms of the sum $\sum_k \mu_k e^{\tau (v^k)' z}$ are those for which $(v^k)' z$ is maximized. That is, by defining the set $I(z) = \{k \in \{1, \ldots, r\} : (v^k)' z = \tilde{Q}(z)\}$ we have that

$$\sum_{k \in \{1, \ldots, r\}} \mu_k e^{\tau (v^k)' z} \approx \sum_{k \in I(z)} \mu_k e^{\tau (v^k)' z} = e^{\tau \tilde{Q}(z)} \sum_{k \in I(z)} \mu_k. \tag{4.9}$$

Given a vector z, define the discrete probability measure $\tilde{\mu}_z$ by

$$\tilde{\mu}_z\{k\} = \begin{cases} \frac{\mu_k}{\sum_{k \in I(z)} \mu_k} & \text{if } k \in I(z) \\ 0 & \text{otherwise.} \end{cases}$$

Using approximation (4.9) in the expressions for the partial derivatives we obtain that

$$\lim_{\tau \to \infty} \frac{\partial \tilde{Q}_\mu^\tau}{\partial z_i}(z) = \mathbb{E}_{\tilde{\mu}_z}\left[\mathbf{v}_i\right] \tag{4.10}$$

$$\lim_{\tau \to \infty} \frac{1}{\tau^{\ell - 1}} \frac{\partial^\ell \tilde{Q}_\mu^\tau}{\partial z_{i_1} \ldots \partial z_{i_\ell}}(z) = \mathbb{E}_{\tilde{\mu}_z}\left[(\mathbf{v}_{i_1} - \mathbb{E}_{\tilde{\mu}_z}[\mathbf{v}_{i_1}]) \ldots (\mathbf{v}_{i_\ell} - \mathbb{E}_{\tilde{\mu}_z}[\mathbf{v}_{i_\ell}])\right]. \tag{4.11}$$

For example, (4.10) says that the limit of the gradients $\{\nabla \tilde{Q}_\mu^\tau(z)\}$ is a convex combination — given by $\tilde{\mu}_z$ — of the gradients of the maximizing functions at z. That is, the limit is some subgradient of $\tilde{Q}$ at z. Conversely, if a vector w is a subgradient of $\tilde{Q}$ at z then there exists a measure μ such that $w = \lim_{\tau \to \infty} \tilde{Q}_\mu^\tau(z)$. Moreover, at points z where $\tilde{Q}$ is differentiable we have that $\tilde{\mu}_z$ has mass one on some k, which implies that $\lim_{\tau \to \infty} \tilde{Q}_\mu^\tau(z) = v^k = \nabla \tilde{Q}(z)$ and the limit of all higher-order partial derivatives is zero (as expected). All these facts are mentioned in Guillaume and Seeger (2001). Note also that the above results generalize those calculated in that paper, where results up to second order were presented.

By considering approximations to $G(x, \boldsymbol{\xi})$ of the form $G^\tau(x, \boldsymbol{\xi}) = c^t x + \tilde{Q}^\tau(h - Tx)$, we see that $G^\tau(x, \cdot)$ is infinitely differentiable and $G^\tau(x, \cdot) \to G(x, \cdot)$ for each x. As discussed in the remark following Assumption D1, that condition — together with the use of an approximation F^Δ for F^{-1} — suffices for (3.16) to hold. Notice also that (4.11) suggests that (3.17) holds as well, at least at the points x such that $\tilde{Q}^\tau(h - Tx)$ is non-differentiable,

a condition that typically holds at the optimal solution x^*. In summary, Assumption D1 holds for approximations G^τ and F^Δ respectively of G and F^{-1}; however, in that case Assumption A3 no longer holds. Moreover, as mentioned before the smoothness comes at the cost of bias of the estimators $\hat{g}_N^{\mathrm{RQMC}}(x)$. Notice also that solving the sample approximation of the problem $\min_{x\in X}\ c^t x+\mathbb{E}[\tilde{Q}^\tau(\boldsymbol{\xi}-Tx)]$ requires different techniques compared to the sample approximation of the original problem — the latter can formulated as a linear program, whereas the smoothed version using $\tilde{Q}^\tau$ requires a nonlinear programming method. Because of that, we did not implement this smoothing technique in the experiments reported in the next section, although we did test the effect of using F^Δ in place of F^{-1}.

5 Numerical experiments

To illustrate the ideas set forth in the previous sections, we discuss now some numerical experiments conducted with two small problems available in the literature. The first problem is `APL1P`, a model for electric power capacity expansion on a transportation network that was first described by Infanger (1992). The second problem is `LandS`, a modification of a simple problem in electrical investment planning originally presented by Louveaux and Smeers (1988). The modified version we study is the one discussed in Linderoth et al. (2005).

APL1P `APL1P` has two decision variables with 2 constraints (plus lower bound constraints) on the first stage, and 9 decision variables with 5 constraints (plus lower bound constraints) on the second stage. The random variables appear on both the right hand side and the technology matrix of the second stage. There are 5 independent random variables. The number of realizations per random variable yields a total of $4\times5\times4\times4\times4 = 1280$ scenarios. With current computing power, this problem can be easily solved exactly; nevertheless, we present the results with sampling because from that perspective the problem is ill-conditioned (cf. Shapiro et al. 2002), which means that the approximating solutions $\hat{x}_N$ are likely to vary with replications. That, in turn, ensures that the objective value estimators $\hat{\nu}_N$ do not correspond to the same solution — if they did, the analysis of rate of convergence would reduce to that of pointwise estimation. Thus, we view this case as a good test for the theoretical results presented in the paper.

We adopted the following methodology. We solved the approximating problem (1.3) using samples generated with standard Monte Carlo, Latin Hypercube Sampling and randomized (t, s)-sequences in base 2 (which, as discussed in section 3.2, is a form of RQMC). For each sampling scheme, we solved the problem with sample sizes equal to successive powers of 2, ranging from 2^5 to 2^{14}. For each sample size, twenty-five replications were run and the standard deviation of the estimators $\hat{\nu}_N$ over these replications was calculated. All simulations used independent random streams. By plotting the logarithm of the standard deviation against the logarithm of the sample size we can visualize the rate of convergence — for example, with standard Monte Carlo one expects to obtain a straight line with slope $-1/2$.

The sampling approximation problems were solved in two steps: first, we used the SUTIL library (Czyzyk, Linderoth, and Shen, 2005) to generate the linear programs corresponding to each sampled problem. SUTIL can construct MPS files for Monte Carlo sampling approximations of two-stage stochastic linear programs; we modified the library slightly to incorporate LHS and randomized (t, s)-sequences, using the publicly available routines developed by Friedel and Keller (2002). The resulting MPS files were fed into the software package Xpress–MP^{TM} from Dash Optimization (under the Academic Partnership Program).

Figure 3 shows the results. We can see that both Monte Carlo and LHS yield a convergence rate of $N^{-1/2}$, thus corroborating the results of Shapiro (1991) for Monte Carlo and of Theorem 3.3 for LHS. The rate for RQMC does not quite follow the result in Theorem 3.4. As remarked earlier (cf. discussion following Theorem 3.4), this is expected due to the discontinuity of the inverse cdf — in this problem the underlying distributions are discrete. Nevertheless, it is clear from the figure that the rate obtained with RQMC is better than both Monte Carlo and LHS. Moreover, the figure shows that both LHS and RQMC yield estimators with smaller variance than Monte Carlo, even though the rate of convergence (in case of LHS) is the same as that of Monte Carlo.

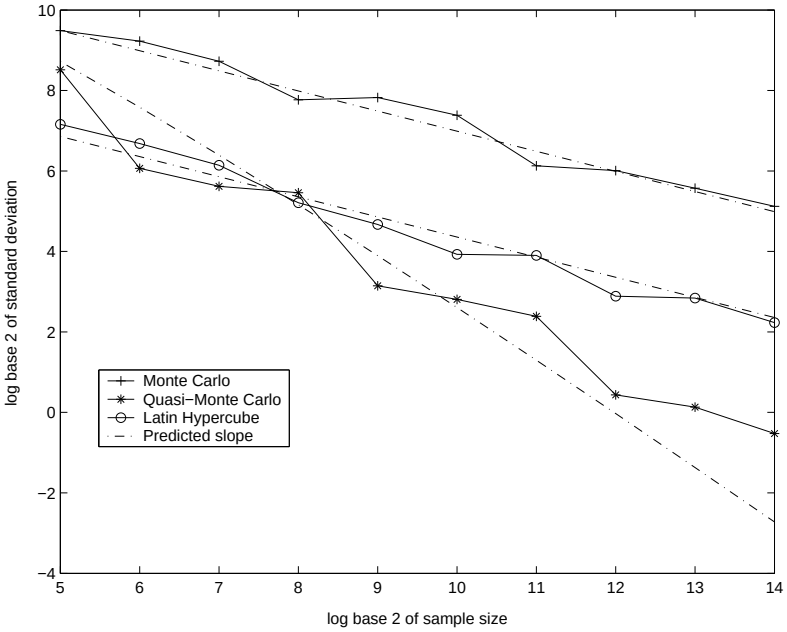

Figure 3: Rates of convergence for `APL1P` problem.

LandS The `LandS` problem has 4 decision variables on the first stage, and 12 decision variables on the second stage. Randomness appears only on the right-hand side of the second stage, in the form of demand constraints. There are three independent random variables, each with 100 possible realizations. Thus, the total number of scenarios is 10^6.

The methodology we adopted was the same as in the `APL1P` case. Figure 4 shows the results. Again, we see that both Monte Carlo and LHS yield a convergence rate of $N^{-1/2}$. The rate for RQMC appears to be of order N^{-1}, which is not as good as the rate in Theorem 3.4

(again due to the discontinuity of F^{-1}) but still better than the Monte Carlo and LHS rates. As in the previous example, we see that both LHS and RQMC yield estimators with smaller variance than Monte Carlo; also, the variance with RQMC is smaller than with LHS.

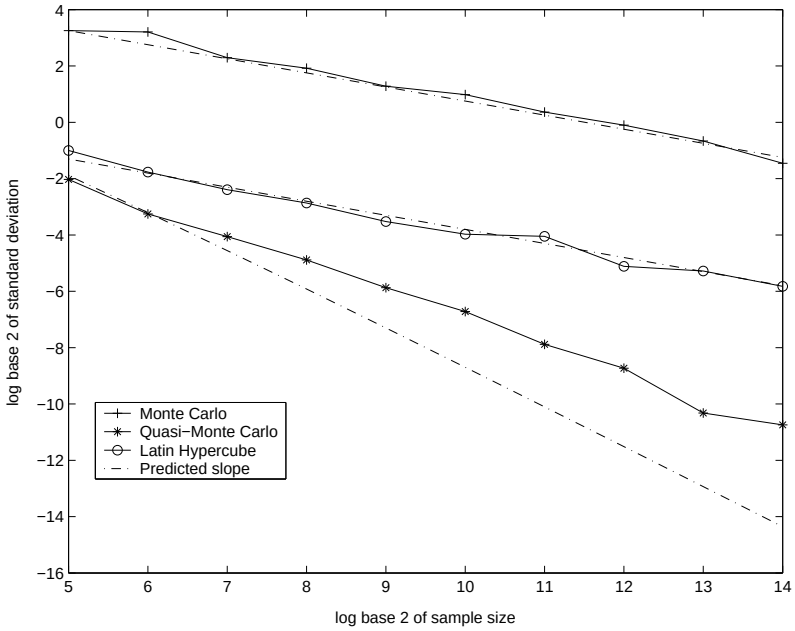

Figure 4: Rates of convergence for `LandS` problem.

Finally, we considered the effect of smoothing the inverse cdfs F_i^{-1} in both examples. We adopted the same technique described in section 3.2 — i.e., F_i^{-1} was replaced by a smooth function F_i^{Δ} such that F_i^{-1} and F_i^{Δ} coincide everywhere except on a interval of size 2Δ around each discontinuity point. Figures 5-8 show the effect of smoothing both on the rate of convergence and on the bias of the estimator $\hat{\nu}_N$. In the `APL1P` problem smoothing helps somewhat, at a cost of a bias of around 0.1% for the higher value of Δ. In the `LandS` problem smoothing works perfectly — with $\Delta = 0.005$ we obtain the rate predicted by Theorem 3.4 without incurring virtually any bias, even though the theorem is not directly applicable in the absence of Assumption A3. This suggests that Theorem 3.4 is valid under more general conditions then those we have used.

6 Conclusions

The theoretical and numerical results in this paper suggest that alternative sampling methods such as Latin Hypercube Sampling and quasi-Monte Carlo can be very effective when solving stochastic optimization problems via sample average approximations. The effectiveness is measured in terms of rates of convergence of estimators of optimal solutions and of optimal values as functions of the sample size. The main contribution of the paper is establishing that rates of convergence for pointwise estimators (i.e. estimation of integrals) carry over to estimators of optimal values/solutions, which allows for the use of results for pointwise estimation available in the literature. In particular, the results in the paper show

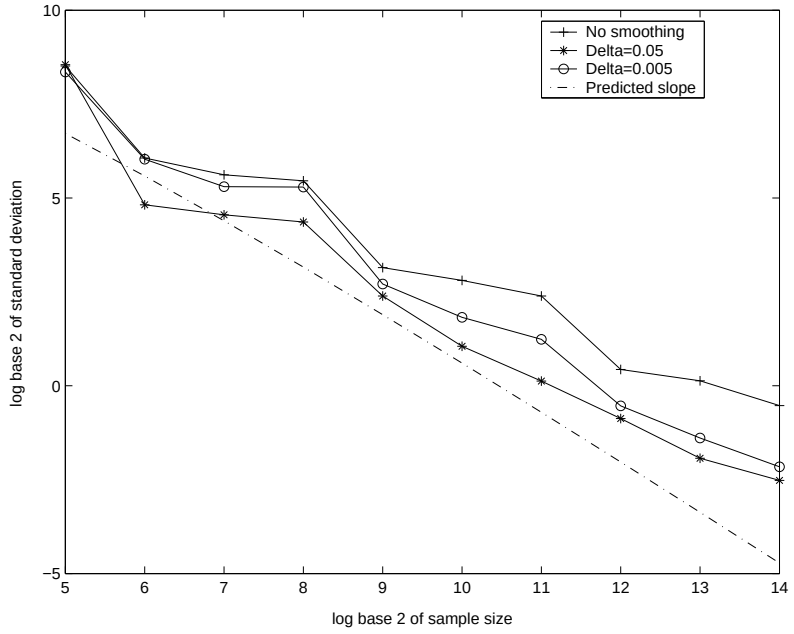

Figure 5: Rates of convergence for `APL1P` problem under RQMC with smoothing.

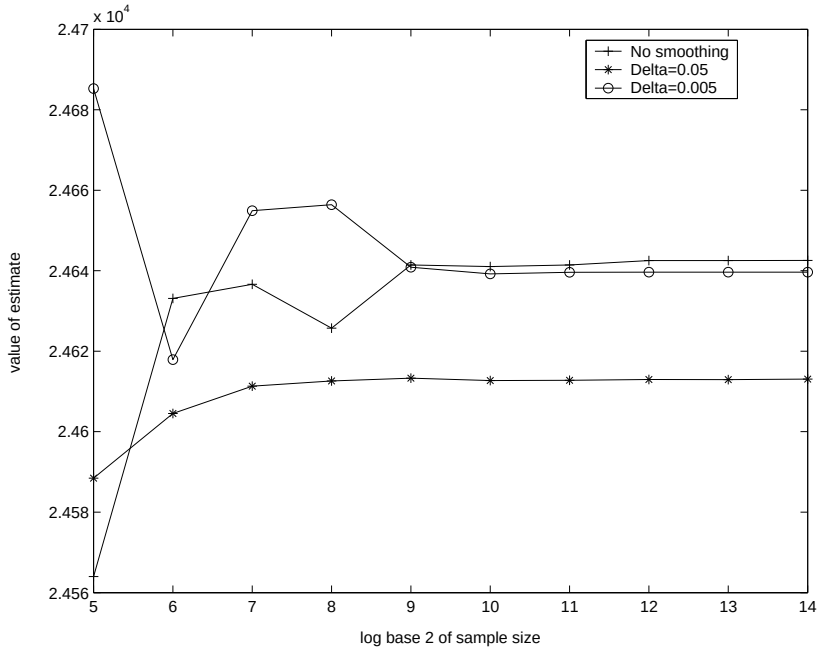

Figure 6: Values of the estimate $\hat{\nu}_N$ for `APL1P` problem under RQMC with smoothing.

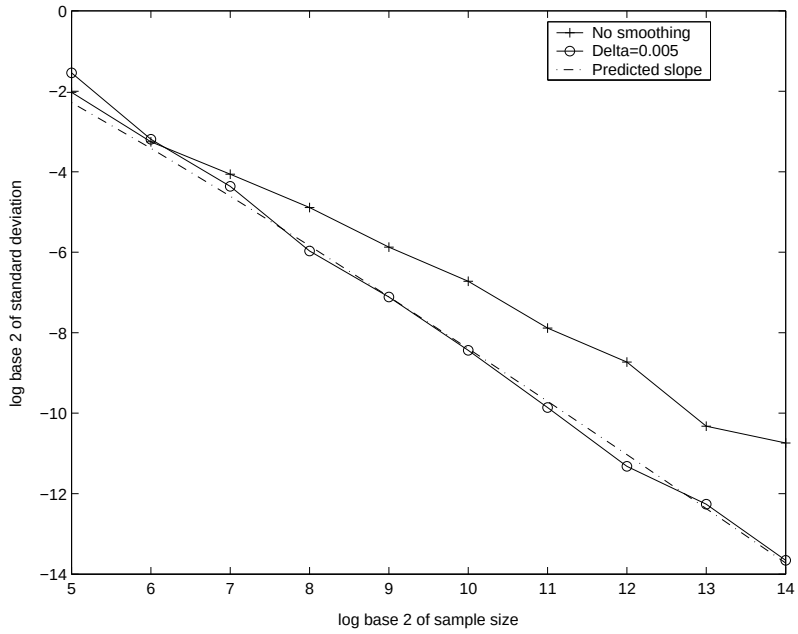

Figure 7: Rates of convergence for `LandS` problem under RQMC with smoothing.

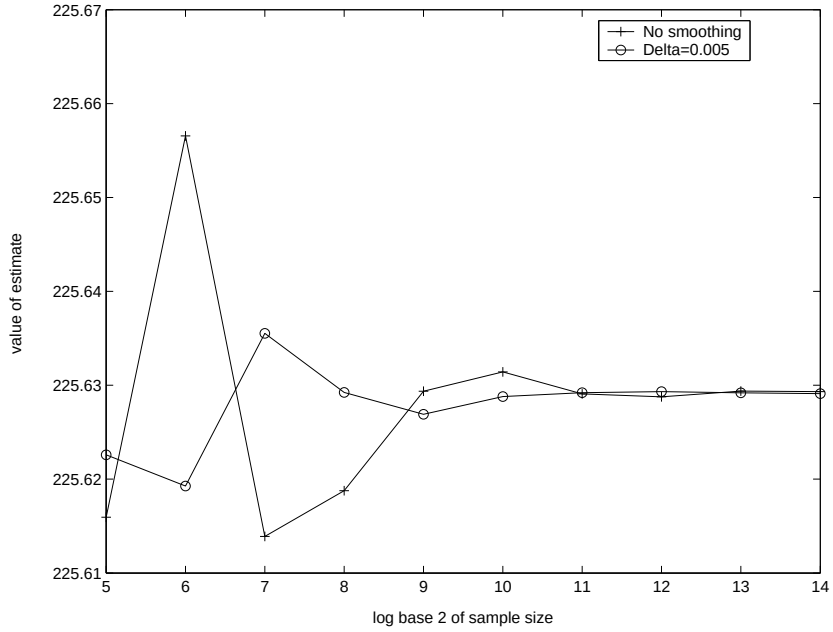

Figure 8: Values of the estimate $\hat{\nu}_N$ for `LandS` problem under RQMC with smoothing.

that, under appropriate conditions, it is possible to obtain a rate of convergence of order $[(\log_b N)^{s-1}/N^3]^{1/2}$ for the approximating optimal values $\hat{\nu}_N$, which asymptotically is much better than the $N^{-1/2}$ obtained with standard Monte Carlo.

Such results are very encouraging, and at the same time raise some interesting issues for further investigation. One topic concerns the effect of smoothing on the rates of convergence when using RQMC — as discussed earlier, the "ideal" rate $[(\log_b N)^{s-1}/N^3]^{1/2}$ derived in Theorem 3.4 seems to require smoothness of the inverse cdf and of the objective function. However, it is unclear whether such conditions are necessary. Our numerical results suggest that smoothness (or at least continuity) of the inverse cdf is crucial, but smoothness of the objective function seems to be less important (cf. the `LandS` example).

On the other hand, our experiments also suggest that Theorem 3.4 is valid even when Assumption A3 does not hold. This is not surprising — as we mentioned earlier, it is possible that a functional version of Assumption B2 holds for the functional space $C(X)$ under RQMC, in which case the conditions of Assumption A3 would not be required; however, we are not aware of the existence of such result.

It would also be interesting to study the precise effect of having multiple optimal solutions on the rates of convergence of optimal values — the main results we have obtained for that case under LHS and RQMC (Theorems 3.3 and 3.4) require uniqueness of the optimal solution. Such a task, however, is likely to require again a functional or at least multivariate version of Assumption B2 (we note that multivariate CLTs have been proved for LHS, but not for RQMC).

Another important issue concerns the dimensionality of the problems. It is well known that the performance of RQMC methods worsens with the number of dimensions — indeed, it is easy to see that, when s is large, the term $[(\log_b N)^{s-1}/N^3]^{1/2}$ only becomes smaller than $N^{-1/2}$ for large N. For example, with $s = 30$ one needs $N \geq 2^{16}$ to get the benefits of the RQMC approach. This suggests that RQMC sampling should be used only with some of the random variables involved in the problem; however, determining which ones to select is a nontrivial issue. Research on this topic is underway.

Acknowledgments

We thank Shane Drew for his help with the numerical experiments, Jeff Linderoth for help with the SUTIL library and Mihai Anitescu for discussions on an early version of this paper. We have also benefited from comments from a group of people who attended our talk at the INFORMS Conference in San Francisco (November 2005), where part of this work was presented. Finally, we thank Alexander Shapiro for his comments on a draft of this paper.

References

T. G. Bailey, P. Jensen, and D. Morton. Response surface analysis of two-stage stochastic linear programming with recourse. *Naval Research Logistics*, 46:753–778, 1999.

R. Bartle. *The Elements of Real Analysis.* Wiley, New York, 2nd. edition, 1987.

P. Billingsley. *Probability and Measure.* Wiley, New York, 3rd. edition, 1995.

J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer Series in Operations Research. Springer-Verlag, New York, NY, 1997.

P. Bratley, B. L. Fox, and L. E. Schrage. *A Guide to Simulation*. Springer-Verlag, New York, NY, 2nd edition, 1987.

K. L. Chung. *A Course in Probability Theory*. Academic Press, New York, NY, 2nd. edition, 1974.

J. Czyzyk, J. Linderoth, and J. Shen. SUTIL: A utility library for handling stochastic programs, 2005. User's Manual. Software available at `http://coral.ie.lehigh.edu/sutil`.

L. Dai, C. H. Chen, and J. R. Birge. Convergence properties of two-stage stochastic programming. *J. Optim. Theory Appl.*, 106(3):489–509, 2000.

G. B. Dantzig and P. W. Glynn. Parallel processors for planning under uncertainty. *Annals of Operations Research*, 22:1–21, 1990.

A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer-Verlag, New York, NY, 2nd. edition, 1998.

S. S. Drew and T. Homem-de-Mello. Some large deviations results for latin hypercube sampling. Manuscript, Department of Industrial Engineering and Management Sciences, Northwestern University, 2005.

J. Dupačová and R. J.-B. Wets. Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. *The Annals of Statistics*, 16:1517–1549, 1988.

M. Emsermann and B. Simon. Improving simulation efficiency with quasi control variates. Manuscript, 2000.

G. Fishman. *Monte Carlo: Concepts, Algorithms and Applications*. Springer-Verlag, New York, NY, 1997.

B. L. Fox. *Strategies for Quasi-Monte Carlo*. Kluwer Academic Publishers, Norwell, MA, 2000.

I. Friedel and A. Keller. Fast generation of randomized low-discrepancy point sets. In *Monte Carlo and quasi-Monte Carlo methods, 2000 (Hong Kong)*, pages 257–273. Springer, Berlin, 2002. Software available at `http://www.multires.caltech.edu/software/libseq/`.

S. Guillaume and A. Seeger. A higher-order smoothing technique for polyhedral convex functions: geometric and probabilistic considerations. *J. Convex Anal.*, 8(1):109–126, 2001.

G. Gürkan, A. Y. Özge, and S. M. Robinson. Sample-path solutions of stochastic variational inequalities. *Mathematical Programming*, 84:313–334, 1999.

J. L. Higle. Variance reduction and objective function evaluation in stochastic linear programs. *INFORMS Journal on Computing*, 10(2):236–247, 1998.

J. L. Higle and S. Sen. Stochastic decomposition: An algorithm for two stage linear programs with recourse. *Mathematics of Operations Research*, 16(3):650–669, 1991.

J. B. Hiriart-Urruty and C. Lemarechal. *Convex Analysis and Minimization Algorithms*, volume I. Springer-Verlag, Berlin, Germany, 1993.

G. Infanger. Monte Carlo (importance) sampling within a Benders decomposition algorithm for stochastic linear programs. *Annals of Operations Research*, 39:69–95, 1992.

G. Infanger. *Planning under Uncertainty: Solving Large Scale Stochastic Linear Programs*. Boyd & Fraser Publishing Company, Massachusetts, 1994.

J. Kalagnanam and U. Diwekar. An efficient sampling technique for off-line quality control. *Technometrics*, 39(3):308–319, 1997.

Y. M. Kaniovski, A. J. King, and R. J.-B. Wets. Probabilistic bounds (via large deviations) for the solutions of stochastic programming problems. *Ann. Oper. Res.*, 56:189–208, 1995.

A. J. King and R. T. Rockafellar. Asymptotic theory for solutions in statistical estimation and stochastic programming. *Mathematics of Operations Research*, 18:148–162, 1993.

A. Kleywegt, A. Shapiro, and T. Homem-de-Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2001.

M. Koivu. Variance reduction in sample approximations of stochastic programs. *Mathematical Programming*, pages 463–485, 2005.

A. M. Law and W. D. Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, New York, NY, 3rd. edition, 2000.

J. T. Linderoth, A. Shapiro, and S. J. Wright. The empirical behavior of sampling methods for stochastic programming. *Annals of Operations Research*, 2005. forthcoming.

W. Loh. On latin hypercube sampling. *The Annals of Statistics*, 24(5):2058–2080, 1996.

W.-L. Loh. On the asymptotic distribution of scrambled net quadrature. *Ann. Statist.*, 31 (4):1282–1324, 2003.

F. Louveaux and Y. Smeers. Optimal investments for electricity generation: A stochastic model and a test problem. In Y. Ermoliev and R. J.-B. Wets, editors, *Numerical techniques for stochastic optimization problems*, pages 445–452. Springer-Verlag, Berlin, 1988.

M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21:239–245, 1979.

H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM, Philadelphia, PA, 1992.

A. B. Owen. A central limit theorem for latin hypercube sampling. *J. Roy. Statist. Soc. Ser. B*, 54:541–551, 1992.

A. B. Owen. Randomly permuted (t, m, s)-nets and (t, s)-sequences. In *Monte Carlo and quasi-Monte Carlo methods in scientific computing (Las Vegas, NV, 1994)*, volume 106 of *Lecture Notes in Statist.*, pages 299–317. Springer, New York, 1995.

A. B. Owen. Monte Carlo variance of scrambled net quadrature. *SIAM J. Numer. Anal.*, 34 (5):1884–1910, 1997a.

A. B. Owen. Scrambled net variance for integrals of smooth functions. *Ann. Statist.*, 25(4): 1541–1562, 1997b.

A. B. Owen. Latin supercube sampling for very high-dimensional simulations. *ACM Transactions on Modeling and Computer Simulation*, 8:71–102, 1998.

A. B. Owen. Monte Carlo, quasi-Monte Carlo, and randomized quasi-Monte Carlo. In *Monte Carlo and quasi-Monte Carlo methods 1998 (Claremont, CA)*, pages 86–97. Springer, Berlin, 2000.

T. Pennanen. Epi-convergent discretizations of multistage stochastic programs. *Mathematics of Operations Research*, 30:245–256, 2005.

T. Pennanen and M. Koivu. Epi-convergent discretizations of stochastic programs via integration quadratures. *Numerische Mathematik*, 100:141–163, 2005.

G. C. Pflug. Scenario estimation and generation. Tutorial presented at the X Stochastic Programming Conference, Tucson, AZ, 2004.

E. L. Plambeck, B. R. Fu, S. M. Robinson, and R. Suri. Sample-path optimization of convex stochastic performance functions. *Mathematical Programming, Series B*, 75:137–176, 1996.

S. M. Robinson. Analysis of sample-path optimization. *Mathematics of Operations Research*, 21:513–528, 1996.

R. Y. Rubinstein and A. Shapiro. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. John Wiley & Sons, Chichester, England, 1993.

A. Shapiro. Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30: 169–186, 1991.

A. Shapiro. Asymptotic behavior of optimal solutions in stochastic programming. *Mathematics of Operations Research*, 18:829–845, 1993.

A. Shapiro. Monte Carlo sampling methods. In A. Ruszczynski and A. Shapiro., editors, *Handbook of Stochastic Optimization*. Elsevier Science Publishers B.V., Amsterdam, Netherlands, 2003.

A. Shapiro. On complexity of multistage stochastic programs. *Operations Research Letters*, 34:1–8, 2006.

A. Shapiro and T. Homem-de-Mello. A simulation-based approach to two-stage stochastic programming with recourse. *Mathematical Programming*, 81:301–325, 1998.

A. Shapiro and T. Homem-de-Mello. On rate of convergence of Monte Carlo approximations of stochastic programs. *SIAM Journal on Optimization*, 11:70–86, 2000.

A. Shapiro, T. Homem-de-Mello, and J. C. Kim. Conditioning of convex piecewise linear stochastic programs. *Mathematical Programming*, 94:1–19, 2002.

M. L. Stein. Large sample properties of simulations using latin hypercube sampling. *Technometrics*, 29:143–151, 1987.